ASIALEX
The Asian Association for Lexicography

KMITL
King Mongkut's Institute of Technology Ladkrabang

The 12th International Conference of the Asian Association for Lexicography

ASIALEX 2018 Full Papers

"Lexicography in the Digital World"
June 8th - 10th, 2018
KRABI, THAILAND

NATIONAL GEOGRAPHIC LEARNING

KDICTIONARIES

SPONSORED BY KMITL

ORGANISED BY FACULTY OF LIBERAL ARTS

ASIALEX 2018

"Lexicography in the Digital World"

Krabi, Thailand
8th – 10th June, 2018

Programme and Abstracts
http://www.kmitl.ac.th/asialex/

Organisers
The Asian Association for Lexicography (ASIALEX)
King Mongkut's Institute of Technology Ladkrabang (KMITL)

WELCOME MESSAGE

On behalf of ASIALEX 2018 Organising Committee, we are delighted to welcome you to the conference in Krabi during the 8th to 10th of June, 2018. We would like to give a Thai traditional greeting 'Sawasdee' to you all of our honoured guests.

The 12th International Conference of the Asian Association for Lexicography (ASIALEX 2018) is supported by King Mongkut's Institute of Technology Ladkrabang, Thailand. This year, the theme of the conference is *Lexicography in the digital World*. We are grateful to have five distinguished keynote speakers: Pedro A. Fuertes-Olivera from University of Valladolid (Spain), Pam Peters from Macquarie University (Australia), John Simpson from University of Oxford (England), Shigeru Yamada from Waseda University (Japan), and Virach Sornlertlamvanich from Thammasart University (Thailand). A total number of 43 papers have been submitted from around the world. All papers have been peer reviewed by Scientific Committee.

Apart from participating in the conference, we invite you to explore the most beautiful beaches and city of Krabi on the west coast of southern Thailand. It offers not only the nature, but also a unique treasure trove of cultural attractions such as temples and traditional ways of life of the southern part of Thailand.

We hope that you have a fabulous experiencing net-working during the conference and enjoy spending time on fantastic white-sand beaches and turquoise water in Krabi, the most beautiful venue in Thailand.

Assoc. Prof. Dr. Jirapa Vitayapirak
Academic Convenour, ASIALEX 2018

CONTENTS

CONFERENCE PROGRAMME

Thursday, 7th June 2018			
14:00-17:00	**Registration and Check-in (ASIALEX Counter near reception at Krabi Resort)**		
Friday, 8 th June 2018			
8:30-9:00	**Opening Ceremony & Group Photo** (Sai Ngoen)		
9:10-9:50	**Keynote Speech I** **Pedro A. Fuertes-Olivera (University of Valladolid, Spain)** **Title: *Designing and Making Commercially-Driven Integrated Dictionary Portals: The Diccionarios Valladolid-UVa*** (Sai Ngoen)		
9:50-10:15	Coffee Break		
Venue	Sai Ngoen	Sai Thong	Watanatam 1
10.15-10.40	*Asian Englishes in the Oxford English Dictionary: Recent Achievements and Future Prospects* **Danica Salazar**	*Collecting Etymological Information of Indonesian Malay Lexicon from Diachronic Corpora* **Dewi Puspita**	*A Comparative Study of Monolingual/ Bilingual Learner's Dictionaries for Encoding Purposes* **Naho Kawamoto**
10.40-11.05	*Translating Definition for Medical Terms in an Age of Collaborative Lexicography* **Jun Ding**	*Building a Corpus-Based Frequency Dictionary for Vietnamese* **Dien Dinh, Triet Nguyen Quang Minh & Thuy Ho Hai**	*Indigenous Languages and Learner Dictionaries: The Use of Javanese Dictionary in University* **Totok Suhardijanto & Atin Fitriana**
11.05-11.30	*Monitoring Academic Studies of Turkish Lexicography: A Photograph of 84 Years* **Ferdi Bozkurt**	*A Corpus-Based Analysis on Lexico-Grammatical Features in Cooking Shows* **Pitchayanin Inla & Kornwipa Poonpon**	*A Discourse Analysis of Editors' Prefaces of Bilingual Dictionaries* **Wai-on Law**

Venue	Sai Ngoen	Sai Thong	Watanatam 1
11.30-11.55	*The Importance of Co- and Context in Solving Lack of Dictionary Equivalence* **Alenka Vrbinc**	*Developing GLOOSH: a Bilingual Digitalized Glossary of Occupational Safety and Health* **Tan Kim Hua**	*Polyonymy on Terminology of Turkish Lexicography* **Fatih DOĞRU**
11.55-13.25	Buffet Lunch (The Boat Restaurant)		
13.25-13.50	**Keynote Speech II** **Pam Peters (Macquarie University, Australia)** **Title: *Issues in Bilingual Terminography for Australian Family Law*** (Sai Ngoen)		
Venue	Sai Ngoen	Sai Thong	Watanatam 1
13.50-14.15	*Translingual Words and Social Media: Are They Asian Words or English Words?* **Brittany Khedun-Burgoine**	*CEFR-Based Grading and Sequencing of Phrasal verbs and its Implications for Pedagogical Lexicography* **Yukio Tono**	*Is the isiNdebele Terminology Developed Today of any Impact?* **K.S Mahlangu**
14.15-14.40	*Pronunciation in EFL Dictionaries: A Case of During in American English* **Kensei Sugayama**	*Towards Building a Language Family Tree for Low-Resource Languages: Clustering Using Orthographic Features* **Angelica Dela Cruz, Nathaniel Oco & Rachel Edita Roxas**	*On the Inclusion of New Words in A New English-Chinese Dictionary* **Gao Yongwei**
14.40-15.05	*A Multi-Dimensional Discrimination of the English Equivalents in Chinese-English Dictionaries for Chinese Users* **Chengmin Liao, Lixin Xia & Mengyu Zhang**	*Integrating Thai WordNet and SenticNet Into Thai Sentiment Resource* **Ponrudee Netisopakul**	*Neologism & Lexicography: Lexicography Challenges in Lemmatizing New Words in the Sotho Languages* **MV Mojela**

Venue	Sai Ngoen	Sai Thong	Watanatam 1
15.05-15.30	*Beyond the Definition: Usage Effects in Dictionaries* **Mehmet Gürlek**	*A Filipino-English Disaster Sentiment Polarity Lexicon* **Joseph Marvin Imperial, Jeyrome Orosco, Shiela Mae Mazo, Lany Maceda, Nathaniel Oco & Rachel Edita Roxas**	*Entryword Choice in Bilingual Dictionaries in the Digital World: New Challenges* **Mats-Peter Sundström**
18:00-19:30	Welcome Reception (Seafood BBQ on the beach at Krabi Resort)		
Saturday, 9th June 2018			
Venue	Sai Ngoen	Sai Thong	Watanatam 1
9:00-9:25	*Beyond a Static Dictionary: Helping Writers with Academic English Collocations* **Robert Lew, Ana Frankenberg-Garcia, Jonathan C. Roberts & Geraint P. Rees Nirwan Sharma**	*Newly Established Idioms Through Blending Semantically Similar Idioms—'Take Care for', 'Take Care about', 'Care of' as Examples* **Ai Inoue**	*Framing Specialized Concepts Through Automatic Extraction and Semantic Annotation: the Case of the DEFORESTATION Event* **Beatriz Sanchez Cárdenas & Carlos Ramisch**
9:25-9:50	*Developing a Finite State Lexicon for Sindhi* **Mutee U Rahman & Hameedullah Kazi**	*The Grammar of Multi-Word Verbs in Philippine English* **Jennibelle R. Ella & Shirley N. Dita**	*The Turkish Lexicography Corpus (TLC): An Overview* **Erdoğan BOZ, Ferdi Bozkurt & Fatih DOĞRU**
9:50-10:15	Tea Break		

10:15-11:55	**Keynote Speech III** **John Simpson (Kellogg College, University of Oxford, United Kingdom)** **Title: *What Did We Think We Were Doing? The Origins of Dictionary Digitalisation*** (Sai Ngoen)		
11:55-12:35	ASIALEX Annual General Meeting (AGM) President's Report		
12:35-14:00	Buffet Lunch (The Boat Restaurant)		
Venue	Sai Ngoen	Sai Thong	Watanatam 1
14:00-14:25	*The Role of Oxford Indonesian Living Dictionary in Conveying Linguistic Knowledge* **Deny A. Kwary**	*A Study on the Use of the Chinese-English Dictionary: What Reference Skills and Strategies are Used by Chinese college Students?* **Mengyu Zhang, Lixin Xia & Chengmin Liao**	*Dictionary and Culture: Lexicography in a Multilingual Context* **Gunter Schaarschmidt**
14:25-14:50	*A Hypergraph Data Model for Building Multilingual Dictionary Applications* **Louis Lecailliez & Mathieu Mangeot**	*Dictionary Look-Up Behavior of Thai engineering Students* **Atipat Boonmoh & Chuthamat Thammajindarach**	*An Investigation Into Students' Perception on Utilizing Online Dictionaries in Translation-Interpretation* **Le, Thi Kieu Van & Dao, Thi Minh Thu**
14:50-15:15	*Gender Orientation in Lexis, Corpora and bilingual dictionaries* **Li Lan & Yue Gu**	*The Effects of Dictionary Use on L2 Error correction* **Yoshiho Satake**	*The Problems in Selecting KBBI's Entry Candidate from Regional Lexicon* **Dewi Khairiah & Dira Hildayani**

15:15-15:40	*Characteristics of Lexical Items in Website Messages of Japanese* **Toshikazu Ezure**	*Using Dictionaries in Metaphor Identification* **Wu Jihong**	*Some Observations About Collocation Dictionary of Adjectives in Turkish* **Bülent Özken**
18:00-19:30	Buffet Dinner (Thai Night at Krabi Resort)		
Sunday, 10th June 2018			
Venue	**Sai Ngoen**	**Sai Thong**	
8:55-9:20	*Chinese Loanwords in English – A Research Based on Oxford English Dictionary* **Shuang Liang**	*Language Documentation and Revitalization: The Case of the Kapampangan Language and its Implication for Mother Tongue-Based Education* **Rosally Viray &** **King Constantine Damaso**	
9:20-10:30	**Panel Discussion** **Virach Sornlertlamvanich (Thammasat University, Thailand)** **Shigeru Yamada (Waseda University, Japan)** **Title: *Lexicography In Asian Context*** (Sai Ngoen)		
10:30-11:00	**Closing Ceremony & Group Photo** (Sai Ngoen)		
11:30-18:00 * 12:00 *	Free Krabi Trip for Participants (Meeting Point: Outside Krabi Resort's Reception) Free Lunch at Roi Thai Restaurant		

EXECUTIVE BOARDS

Rachel Edita O. ROXAS, President
National University (The Philippines)
Vincent BYOOI, Vice-President
National University of Singapore (Singapore)
Shirley DITA, Secretary
De La Salle University (The Philippines)
Deny KWARY, Treasurer
Universitas Airlangga (Indonesia)
Yongwei GAO, Board member
Fudan University (China)
Lan LI, Board member
The Hong Kong Polytechnic University (Hong Kong)
Yukio TONO, Board member
Tokyo University of Foreign Studies (Japan)
Jirapa VITAYAPIRAK, Convenour, ASIALEX2018
King Mongkut's Institute of Technology Ladkrabang (Thailand)
Mehmet GURLEK, Convener, ASIALEX2019
Istanbul University (Turkey)
Hai XU, Co-Chief Editor of LEXICOGRAPHY
Guangdong University of Foreign Studies (China)
Shigeru YAMADA, Co-chief Editor of LEXICOGRAPHY
Waseda University (Japan)
Ilan KERNERMAN, Past President
K Dictionaries (Israel)

REVIEWERS

Assoc. Prof. Dr. Daniel Rueckert,
California State University Fullerton, USA

Assoc. Prof. Dr. Jirapa Vitayapirak,
King Mongkut's Institute of Technology Ladkrabang, Thailand

Assoc. Prof. Dr. Lan Li,
The Chinese University of Hong Kong,Shenzhen, China

Assoc. Prof. Dr. Shirley Dita,
De La Salle University, The Philippines

Assoc. Prof. Dr. Vincent B Y Ooi,
National University of Singapore, Singapore

Assoc. Prof. Dr. Yukio Tono,
Tokyo University of Foreign Studies, Japan

Asst. Prof. Dr. Anongnad Petchprasert,
King Mongkut's Institute of Technology Ladkrabang, Thailand

Asst. Prof. Dr. Atipat Boonmoh,
King Mongkut's University of Technology Thonburi, Bangkok

Asst. Prof. Dr. Passapong Sripicharn,
Thammasat University, Thailand

Asst. Prof. Dr. Kornwipa Poonpon,
Khon Kaen University, Thailand

Dr. Deny A. Kwary,
Universitas Airlangga, Indonesia

Dr. Katarzyna Ancuta,
King Mongkut's Institute of Technology Ladkrabang, Thailand

ORGANISING COMMITTEE

Assoc. Prof. Dr. Jirapa Vitayapirak

Asst. Dr. Anongnad Petchprasert

Dr. Montha Polrak

Dr. Katharzyna Ancuta

Mr. Woraprat Manowang

Mr. Choedphong Uttama

Mr. Patipan Bandurat

Mrs. Pranee Nilkhao

Miss Rattana Sangchan

Miss Suttapa Chanplang

Mr. Tharathep Rattanawan

FULL PAPERS

Keynotes

Pedro A. Fuertes Olivera

Pedro Olivera is currently working as Full Professor at the University of Valladolid in Spain and has been Extraordinary Professor at the Department of Afrikaans and Dutch, University of Stellenbosh in South Africa since 2014. His research interest lies in lexicography, especially in specialised lexicography. In particular, he has employed a functional approach to dictionary making and criticism. Pedro Olivera has been part of the editor team of the *Routledge Handbook of Lexicography*, which was published in October 2017. He has been a keynote speaker at various conferences.

For more details, see http://www.pedrofuertes.net/ and often referred for the journals such as *Lingua*, *Journal of Pragmatics*, *Applied Linguistics*, *English for Specific Purposes* and *International Journal of Lexicography*.

Designing and Making Commercially-Driven Integrated Dictionary Portals: The *Diccionarios Valladolid-UVa*

Pedro A. Fuertes-Olivera
International Centre for Lexicography (Universidad de Valladolid, Spain)
Department of Afrikaans and Dutch (University of Stellenbosch, South Africa)
pedro@emp.uva.es
pedro.a.fuertes@gmail.com

Abstract

Engelberg & Müller-Spitzer (2013: 1023) define a dictionary portal as "a data structure that is presented as a page or set of interlinked pages on a computer screen and provides access to a set of electronic dictionaries, and where these dictionaries can also be consulted as standalone products". Based on several criteria –type of access provided, cross-reference structures, ownership relations between the portal and the dictionaries, layout of the portal – they propose a typology of dictionary portals comprising (a) dictionary nets, (b) dictionary search engines, and (c) dictionary collections. Their analysis as well as that of Boelhouwer, Dykstra and Sijens (2018) conclude that dictionary portals are widespread and that they may illustrate a way ahead for the future of e-lexicography. My view is that dictionary portals can become a solution of commercially-driven lexicography− i.e. the design, compilation and updating of dictionaries with the aim of making an economic profit − if they are conceptualized as lexicographic reference services subjected to efficient cost-benefit analyses. The *Diccionarios Valladolid-UVa* is our answer to this challenge. It is a lexicographic project developed by the joint efforts of lexicographers, IT and domain experts as well as companies. It was started in the year 2014, initially following Bergenholtz's works in several commercially-oriented Danish dictionaries. The project has now been accommodated to new lexicographic and technological developments, has used around 500,000 euros of public and private funds in its development and construction, and is on the road to "producing" 28 *integrated* dictionaries −12 of them deal with English and Spanish accounting language, facts and things; 7 of them mostly describe general Spanish; and 9 of them focus on English and Spanish general language, facts and things −. This project is an *integrated dictionary portal*, i.e. "a tool composed of several three-component devices: (a) editors; (b) search engines; (c) dictionary interfaces. The editors and interfaces of each device are connected by means of search engines equipped with technologies for retrieving dynamic dictionary articles, i.e. different lexicographic data for each user situation. Dictionary articles are prepared by the same team with the basic aim of helping human and/or machine users in several

type situations. They contain both lexicographically prepared data and open linked data with lexicographic value. This lexicographic data can be transferred among devices and adapted, if needed. And this tool can be accessed around the clock either freely or by paying a subscription fee" (Fuertes-Olivera, 2016). This talk will present the philosophy underlying the project as well as some of the technologies and methodologies used in its design, construction and around-the-clock updating (the dictionary articles are available at the moment they are created).

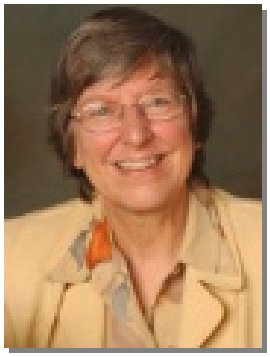
References

Boelhouwer, Bob, Dykstra. Anne and Sijens, Hindrik (2018): "Dictionary portals", in Pedro A. Fuertes-Olivera (ed.), *The Routledge Handbook of Lexicography*, pp. 754-66. London and New York: Routledge.

Engelberg, J. and Müller-Spitzer, C. (2013): "Dictionary Portals", in R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds.): *Dictionaries.An International Encyclopedia of Lexicography*, pp. 1023-1035. Berlin: De Gruyter Mouton.

Fuertes-Olivera, Pedro A. (2016): "European Lexicography in the Era of the Internet: Present Situations and Future Trends", plenary talk, Beijing , 2 December, 2016. Talk sponsored by the Commercial Press and the Chinese Association of Lexicography.

Pamela Hardy Peters



Pam Peters was Director of the Dictionary Research Centre at Macquarie University (2001-7) and a member of the Editorial Committee of the Macquarie Dictionary for its second, third and fourth editions (1991, 1997, 2005, respectively). She was appointed Emeritus Professor on her retirement in December 2007. Her research interest concerns with lexicography, lexicology, terminology and terminography. Her recent work includes *Cambridge Dictionary of English Grammar* (2013), *Cambridge Guide to English Usage Chinese edition* (2011) and *Comparative Studies in Australian and New Zealand English: Grammar and Beyond* (2009).

Bilingual Terminography for Australian Family Law

Pamela Hardy Peters
Macquarie University, Australia
Pam.peters@mq.edu.au

Abstract

While the focus on users has become central to the design of dictionaries, its applications in specialised lexicography and terminography, and especially in online dictionaries and termbanks, are still evolving in their infinite variety in print and online. This presentation examines how the user focus helps to frame the design of an online termbank in law, in an Australian venture in socioterminology.

Every aspect of the *LawTermFinder* termbank in Family Law is motivated by its intended users, they being members of the Australian public who need to understand legal terms that impact on their family relationships. This makes them a special category of learner, neither experts nor trainees in the special subject, but lay people involved in legal actions, seeking very particular legal knowledge to understand their situation. This practical need means that the selection of terms for the termbank's macrostructure is not purely those belonging to family law itself, but others relating to Australian legal systems and processes. Since mediation is a preliminary to commencing divorce proceedings, terms relating to the mediation system must be included.

The microstructure of the termbank is designed in line with the fact that its users – members of the general public seeking family law information – may not be either highly educated or literate in English. The definitions of terms must therefore be in accessible English, and they are accompanied by audio-recordings to support users with low levels of literacy. Where possible, diagrams and tables are used to provide alternative paths to understanding, and to illustrate the relationships among sets and clusters of terms so they do not have to be learned in isolation. This use of multimedia provides enriched contexts for acquiring new terminology, in line with good pedagogical practice for language acquisition. It also serves the needs of termbank users whose first

language is not English – up to a quarter of the current Australian population. The termbank provides translations into 7 community languages of key elements in on each termpage. The selection of these out of the more than 150 immigrant languages spoken in Australia is again decided on the basis of users' needs, both the size of their community, and their self-professed levels of proficiency in English, as summarized in the Australian census every five years.

The translation of Australian law terms into languages other than English confronts us with numerous cultural differences, in the articulation of law as well as the anisomorphism between languages in crucial areas such as referencing family members. In bilingualising LawTermFinder for L2 users of English in Australia, we are also bridging the gap between two legal systems, so that the termbank users can understand the terms in the Australian context.

In all these ways, *LawTermFinder* is an initiative in socioterminology as well as descriptive terminography. By illuminating the meanings of Australian legal terms with verbal, audio and graphic means, as well as bilingualisations, we provide fuller access to their meanings for a spectrum of novice users, whether English is their first or second language.

John Simpson

John Simpson received a BA in English Literature at the University of York in 1975 and an MA in Medieval Studies at the University of Reading in 1976. He has been awarded honorary D.Litts. by the Australian National University and the University of Leicester for his work as a lexicographer. He was appointed as Chief Editor of the Oxford English Dictionary (OED) in 1993 and has worked for Oxford Dictionary until 2013. John Simpson is also a member of the English Faculty and Emeritus Fellow of Kellogg College, University of Oxford. John Simpson's editorship covers the Concise Oxford Dictionary of Proverbs (1982) and co-edited the Oxford Dictionary of Modern Slang (1992). In an international context, he has acted as adviser to a number of national dictionaries, and in 1999 he was awarded an honorary degree by the Australian National University for his distinguished creative achievement as a scholar in lexicography.

Attached is the link of John Simpson talking about how dictionaries changed https://www.youtube.com/watch?v=EwDuJ_gnKh0

What Did We Think We Were Doing? The Origins of Dictionary Digitalisation

John Andrew Simpson
Kellogg College, University of Oxford (United Kingdom)
john.simpson@ell.ox.ac.uk

Abstract

There was a time before we expected dictionaries to be digital. Back then, dictionaries were consulted to discover information about individual words. But in the 1970s and 80s things started to change and since then English dictionaries have repurposed themselves for the digital age. The Oxford English Dictionary was in the forefront of this tide of change, but what did it mean for the editors and for the words themselves? As well as new working methods, there were new aspirations: online dictionaries could report on changing patterns of language, and hence on social and cultural change as documented by the language. But aspiration is not achievements: just how far along the line are we in the history of the digital dictionary today?

Virach Sornlertlamvanich

Virach Sornlertlamvanich earned D.Eng. in Computer Science at Tokyo Institute of

Technology in Japan. At the moment, Virach works as a lecturer at Sirindhorn International Institute of Technology, Thammasat University in Thailand. His research interest lies in natural language processing (NLP), machine translation, and corpus-based approach NLP. His recent work is LEXiTRON, a widely used application of electronic Thai-English dictionary released in 1995 and Thai Royal Institute Dictionary (TRID).

For his complete biography, visit https://www.siit.tu.ac.th/personnel.php?id=139. See https://scholar.google.com/citations?user=HpDlacUAAAAJ&hl=en
for more information on his full research.

From LEX*i*TRON to Asian WordNet
Issues in Language Resource Development

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology, Thammasat University, Thailand
virach@siit.tu.ac.th

Abstract

Several approaches have been studied to cope with the exceptional features of non-segmenting languages. When there is no explicit information about the boundary of a word, segmenting an input text is a formidable task in language processing. Not only the contemporary word list, but also usages of the words have to be maintained to cover the use in the current texts. The accuracy and efficiency in higher processing do heavily rely on this word boundary identification task. Therefore, we introduce some statistical based approaches to tackle the problem due to the ambiguity in word segmentation. The word boundary identification problem is then defined as a part of others for performing the unified language processing in total. To exhibit the ability in conducting the unified language processing, we selectively study the tasks of language identification, word extraction, dictionary-less search engine and term-based ontology alignment. We firstly applied the proposed statistical based approaches to extract the term candidates and their usage examples from the Internet available webpages to create the core of LEX*i*TRON, to be the first Thai-English corpus based dictionary. The development of LEX*i*TRON is continuously conducted by voluntarily contribution up to the present day. Extending the terminology to the ontology, Princeton WordNet (PWN) with its expression power of senses in terms of synset (a set of synonyms), it can facilitate the computational expression very well. We adopted the advantages of sense expression by a list of words, the so call synset, to provide a common platform for collaborative WordNet construction for a language. To prepare an initial WordNet for a certain language, we align the synset to a list of words from the existing bi-lingual dictionaries. The degree of confidence in the synset assignment has been proposed by computing the distance between a word to a member of a synset. Word synonyms are also used to serve in finding a candidate of synset. As a result, the list of candidate synsets is proposed to a word entry together with a degree of confidence score. We have shown the efficiency in nominating the synset candidate by using the most common lexical information. The approach has been successfully evaluated in aligning the terms from several Asian languages to PWN, to form the Asian WordNet (AWN). Currently, the AWN is aligned to be a part of Multilingual WordNet (MWN).

Shigeru Yamada

Shigeru Yamada is currently working as Professor at the Faculty of Commerce, Waseda University in Japan. His research interest focuses on EFL lexicography, bilingual lexicography and teaching of dictionary use. His recent work consists of The L1 Guide to the Usage of an EFL Dictionary: the Case of the Japanese Guide to the Usage of Oxford Wordpower Dictionary, Dictionary Use in Urban Society: Web-based and Hand-held Electronic Dictionary and Monolingual Learners' Dictionaries – Where's now published in The *Bloomsbury Companion to Lexicography*. Further, Shigeru has made contributions to the community of Asian lexicography as Co-chief Editor of Lexicography and Treasurer. For detailed information, consult with http://researchers.waseda.jp/profile/en.a47729b452f20aa129ab04cfc095ea00.html

The Interaction Between EFL and English-Japanese Dictionaries

Shigeru Yamada
Waseda University, Japan
shayamda@waseda.jp

Abstract

EFL dictionaries and English-Japanese dictionaries (EJDs) have developed through the interaction with each other. The grading of headwords was initiated by the *Standard EJD* on the basis of Thorndike's and Horn's wordlists and was refined by EFL dictionaries using corpus data. EJDs preceded EFL dictionaries in the indication of stress patters of compounds. Plausibly, the diagram showing sense development in the *Lighthouse EJD 1* inspired the *Macmillan ED*'s provision of menus. *A Grammar of English Words*' indication of verb patterns (through the *Idiomatic and Syntactic ED*) influenced both EFL dictionaries (initially using codes) and EJDs (transparent indications). *Saito's Idiomological EJD* explicitly indicated the combination of selectional restrictions and verb patterns. Corpus-based lexicography has spread from the *COBUILD1* to other EFL dictionaries and EJDs. Although there may have been antecedents and the input from other sources, the interaction between the two genres of learners' dictionaries enhanced their mutual development.

Keywords: corpus, EFL dictionary, English-Japanese dictionary, grading of headwords, menu, selectional restriction, signpost, stress pattern, verb pattern

The Importance of co- and Context in Solving Lack of Dictionary Equivalence

Alenka Vrbinc
University of Ljubljana
alenka.vrbinc@ef.uni-lj.si

Abstract

The paper deals with the treatment of absence of dictionary equivalents at word level in a decoding English-Slovene dictionary (ESD). Attention is paid to those cases which are solved neither by a descriptive equivalent nor by a loan word. The study whose findings are presented in this contribution focuses on the use of the hash sign, which implies no equivalence at the word level; however, if the untranslatable SL lexical item is used in an example illustrating its use, it can be rendered into the TL, which means that equivalence is reached at the level of the entire message. The entire dictionary was taken as a base for extracting cases of this type of equivalence. A hash sign can be found to mark the absence of equivalence in one or exceptionally several senses of the lemma or phraseological unit. Detailed results are presented by parts of speech of the lemmata, which is followed by analysis of the content of illustrative examples. Then follows a detailed discussion of the lemmata that express pragmatic meaning in the SL, the number of SL senses included under one sense with a hash in the ESD and examples combining lexical and grammatical characteristics. One of the most important conclusions is that, if equivalence cannot be achieved at word level but is possible at the level of the entire message, the problem can be resolved by including translated examples of use. This can be considered appropriate if the dictionary is to function as an effective communication tool.

Keywords: bilingual dictionary, zero equivalence, examples of use, co-text, context

1. Introduction

The most salient element of a lexicographic description is the semantic component. Research into dictionary use and users' reference needs has shown that finding out the meanings of lexical items occupies first place among the reasons for consulting a dictionary. The semantic part of the dictionary entry in monolingual dictionaries is represented by a definition, while in bilingual dictionaries, the dictionary equivalent is provided. Since language learners still often show a preference for bilingual dictionaries, special attention should be given to dictionary equivalence in bilingual dictionaries.

Bilingual lexicographers are expected to find equivalents in the target language (TL) that correspond semantically to the source language (SL) lexical items not only in one particular context but more universally (Adamska-Sałaciak 2010: 388; Atkins 1992/1993: 44f). Lexicographers, however, often come across cases when they fail to find suitable equivalents. The provision of dictionary equivalents in the TL often depends on co-text or context. Consequently, carefully selected co-text or context provided in a bilingual dictionary

in the form of illustrative examples plays a very important role, since examples disambiguate and/or specify the meaning of the lexical item in question (Zgusta 1971: 337).

The relation between the SL lexical item and the TL lexical item is regarded as the equivalent relation (Gouws 2002: 195–196). Equivalent relations are generally of three types, which are referred to as full, partial and zero equivalence (Zgusta 1971: 312–325; Wiegand 2002; Gouws 2002: 196). Full equivalence implies that the SL and the TL lexical items are equivalent lexically, pragmatically and semantically. Full equivalents are quite rare, as they must fit into all contexts and must agree with the SL lexical item not only in denotation but also in connotation (Zgusta 1971: 312; Gouws 2002: 196). The equivalent relation that is most common in bilingual dictionaries is partial equivalence: the semantic component of the dictionary entry consists of several TL equivalents that cover the entire spectrum of meaning of the SL item (cf. Zgusta 1971: 315). Zero equivalence is characterized by a lack of equivalent in the TL and is solved by means of explanatory or descriptive equivalents, loan words or brief paraphrases (cf. Zgusta 1971: 319; Gouws 2002: 200).

This paper examines the treatment of absence of dictionary equivalents in a decoding English-Slovene dictionary (hereafter referred to as ESD) which is in its final stages of completion. In the ESD, the two symbols are employed to mark the absence of equivalents in the TL:

- the slashed zero (Ø) indicates a complete absence of any equivalent;
- the hash sign (#) implies no equivalence at the word level, but if the untranslatable SL lexical item is used in an illustrative example, it can be rendered into the TL, which means that equivalence is reached at the level of the entire message.

This clearly indicates the importance of co-text and context in retrieving semantic information on the SL lexical items. In this paper, the focus is on the use of the hash sign only, and individual cases marked by this symbol in the ESD are addressed in more detail.

2. Methodology

The ESD contains about 53,000 lemmata and about 16,000 secondary lemmata and is the only pedagogically- and didactically-oriented bilingual dictionary compiled so far in Slovenia. Since the study focused on zero equivalence at word level marked with a hash sign (#), the entire dictionary was taken as a base for extracting cases of this type of equivalence. Altogether, 41 lemmata were extracted, but in different sections of one dictionary entry up to three cases of zero equivalence at word level can be identified; therefore, the total number of hash signs used in the ESD is 92. In these cases, semantic information is provided by means of examples of use. The material collected was first analysed as to the part-of-speech of the lemmata to see whether any part of speech stands out as regards zero equivalence at word level. Each sense marked with the hash was then investigated more thoroughly and special attention was paid to examples to see whether these examples share any features that could potentially be regarded as a reason for zero equivalence at word level.

3. Dictionary entries with the hash sign

A careful look at the parts of speech of the lemmata reveals that hash signs can be found in 12 verbal lemmata, 11 nominal lemmata, 7 adjectival and 5 prepositional lemmata. A hash sign can be found much less frequently in other parts of speech (3 lemmata).

3.1 Verbal lemmata

The majority of verbal entries contain a hash sign to indicate lack of dictionary equivalent in one sense of the lemma, the exception being the verb *come* with three such senses, i.e., 8, 9 and 12. Senses 8 and 9 additionally include an explanatory phrase or a pattern illustration:

- sense 8: the explanatory phrase *v vprašalnih stavkih za how* 'in questions after how' precedes the hash, which is followed by translated examples (e.g., *How do you come to be so late?* Kako to, da si tako pozen?);
- sense 9: *come sth (with sb)* before the hash signifies a pattern illustrated by translated examples (e.g., *Don't come the innocent with me!* Ne delaj se nedolžnega!);
- sense 12: translated examples with the following two patterns: *come* + gerund (e.g., *come flying* prileteti) and *come* + prepositional phrase (*come into effect* stopiti v veljavo, začeti veljati).

In the verbal lemma *go*, a hash is used in the idioms section to mark the absence of a dictionary equivalent for the phraseological unit *be going to do sth* which is translated into Slovene grammatically by the future tense form (*She's going to ring us.* Poklicala nas bo.). Apart from that, the phraseological unit is accompanied by the explanatory phrase *za izražanje prihodnosti* 'used to express future'. In the ESD as well as in monolingual learner's dictionaries, *be going to do sth* can be found in the idioms section, but a non-native speaker of English is unlikely to search the idioms section for this word combination. In the process of learning English as a foreign language, learners are taught that this structure is used to form the future tense. Therefore, it would be advisable to treat this structure as a separate sense with a fixed pattern and a brief theoretical explanation followed by illustrative examples.

If we study the content of the examples illustrating the senses marked with a hash, we can see that the lemma used in these examples can be explained in a monolingual English dictionary by means of a pragmatic definition, e.g., *bless* is defined as 'used to express surprise'. In the ESD, a hash is used in place of a dictionary equivalent and the use of *bless* is illustrated by examples following the pattern *bless sb/sth* (*Bless me!* Za božjo voljo!) or containing *be blessed (if ...)* (*I'm blessed if I know!* Naj me vrag, če vem.).

A semantic connection between the verb used in the Slovene translation of the English examples and the noun used in the English example can be observed quite frequently. The sense of the verb *bear* marked with the hash is illustrated by examples of the type *bear* + noun (e.g., *bear a grudge against sb* zameriti komu). All examples are translated into Slovene by a verb (e.g., *zameriti*) whose meaning is related to that of the English noun (e.g., *grudge* = *zamera*).

3.2 Nominal lemmata

Brainchild is a monosemous noun, which means that a dictionary user is offered no equivalents in the TL and can infer the meaning of the lemma from one translated example only (*The system was his own brainchild.* On je bil duhovni oče tega sistema.).

In the entries for some nouns (i.e., *accident, amount* and *comfort*), the hash appears in the idioms section to mark a lack of dictionary equivalents for a phraseological unit (i.e., *an accident of birth; too close/near for comfort; no amount of sth will do sth*). The meaning of *too close/near for comfort* is illustrated by the example *The bombs fell in the sea, many too close for comfort.* Bombe so padale v morje, mnoge veliko preblizu. The translation into Slovene depends on the English adverb (*close, near*) used in the phraseological unit (*too close/near = preblizu*).

In the entry for *animal*, dictionary equivalents cannot be provided for the sense 'a particular type of person, thing, organization, etc.'; the translation of the examples into Slovene (e.g., *That's a queer sort of animal.* To je čuden tič.) depends on the modifiers describing the type of person, thing, organization, etc. (*queer = čuden*) rather than on the noun *animal*.

Pragmatic meaning can be observed in the entry for the noun *goodness* defined as 'used to express surprise'. In this sense, the noun lacks dictionary equivalents in Slovene, but the illustrative examples show that the noun is used in more or less fixed expressions. The translation into Slovene clearly reflects the pragmatic meaning of the English noun (*My goodness!* or *Goodness me!* or *Goodness gracious (me)!* Moj bog!).

3.3 Adjectival lemmata

The adjective *delayed-action* is monosemous, which means that the hash sign indicates a complete lack of dictionary equivalents. The examples follow the pattern *delayed-action* + noun and could be regarded as compounds, or more precisely, as terms; they are also rendered into Slovene as such: e.g., *delayed-action mechanism* samosprožilec.

Three adjectival lemmata marked with a hash in the ESD, i.e., *gracious, great* and *holy*, have the following three characteristics in common:

- The sense with no dictionary equivalents in Slovene has a pragmatic definition in English, e.g., *great* 'used to express shock or surprise'.
- They express restrictions and constraints regarding usage, which are reflected in the accompanying labels, e.g., *great* labelled *spoken old-fashioned*.
- The examples of use are fixed expressions, e.g., *Great heavens!* Za božjo voljo.

3.4 Prepositional lemmata

In prepositional lemmata, the sense marked by a hash in the ESD contains examples illustrating different senses of the lemma in English. For example, the entry for *by*:

- 'used to say that something happens in a particular kind of light': e.g., *by day* podnevi; *by daylight* pri dnevni svetlobi;
- 'used to state the rate at which something happens': e.g., *day by day* iz dneva v dan; *bit by bit* pomalem;
- 'used before particular nouns without *the*, to say that something happens as a result of something': e.g., *by mistake* pomotoma; *by accident* po nesreči;
- 'used to show how something is done': e.g., *by yourself* sam.

With the exception of *by yourself*, which is translated by the pronoun *sam*, all other examples are rendered into Slovene either by an adverb (e.g., *podnevi, pomalem, pomotoma*) or by a prepositional phrase introduced by different prepositions (e.g., *pri dnevni svetlobi, iz dneva v dan, po nesreči*). Such an abundance of translation options results in a great number of examples illustrating the sense with a hash in the ESD.

3.5 Adverbial and pronominal lemmata and lemmata without part-of-speech label

In the entry for the adverb *jolly*, the hash appears in the idioms section to mark the absence of dictionary equivalents for the phraseological unit *jolly well* (e.g., *I'm going to jolly well tell him what I think of him!* Mu bom že povedal, kaj si mislim o njem.).

The only pronominal lemma with a hash sign is *one*. The hash is used to indicate the absence of equivalents in the sense defined as 'used to avoid repeating a noun, when you are

referring to somebody/something that has already been mentioned, or that the person you are speaking to knows about'. *One* in the illustrative examples can be rendered into Slovene by using a substantivized adjective (i.e., *dragi* in *her loved ones* njeni dragi) or a pronoun (i.e., *jo* in *I'd like a cup of tea. Are you having one, too?* Rada bi skodelico čaja. Ali bi jo ti tudi?).

In the ESD, three lemmata with a hash sign lack the part-of-speech label, i.e., *d'you, gonna* and *let's*. All three lemmata are monosemous and represent short or informal forms. The example with *d'you* is translated as a question in the present and past tense forms (*What'd you say?* Kaj praviš?, Kaj si rekel?), *gonna* is translated using the future tense form in Slovene (*This isn't gonna be difficult.* To ne bo težko.) and *let's* is translated using the imperative form of the verb in the first person plural (e.g., *Let's go to the cinema.* Pojdimo v kino.). All three lemmata are translated grammatically rather than lexically. Apart from that, the use of *gonna* is additionally described by a short explanatory phrase, i.e., pogovorna oblika za *going to*, ki izraža prihodnost 'an informal form for *going to* used in reference to the future'.

4. Discussion

The absence of dictionary equivalents should be adequately addressed by bilingual lexicographers, who should base their decisions and solutions on an in-depth lexical contrastive analysis. Special attention should be paid to the various co-texts and contexts in which the lexical item in question is used in the TL and to its rendering into the SL. If different co-texts and contexts suggest that the provision of a dictionary equivalent is not possible, this does not necessarily mean that the lexical item in the SL is untranslatable or untranslated: its equivalence can only be observed at the level of the entire message rather than at the word level. Consequently, the only possibility to treat such lexical items in a bilingual dictionary is to include examples of use. The next issue that should be addressed is whether or not illustrative examples included in a bilingual dictionary should be translated. Metalexicographers have conflicting opinions regarding this matter. Some advocate that the examples should be translated (Al-Kasimi 1977: 96; Zöfgen 1991: 2898), whereas the others believe that there is no need to translate examples if they are chosen so as to pose no problems for anyone with a basic knowledge of the TL (Jacobsen et al. 1991: 2786; Adamska-Sałaciak 2006: 493–494). In the case of zero equivalence, untranslated examples are of no help to the dictionary user, who is not familiar with the meaning of the lemma in question; therefore, the translation of examples is an absolute must (Vrbinc and Vrbinc 2016: 308).

The results of our study show that there is a notable lack of dictionary equivalence in those senses of the lemmata that express pragmatic meaning in the SL. Among the lemmata commonly defined by means of pragmatic definitions, grammatical words, interjections or pragmatic phraseological units should be enumerated, since they have evolved meanings that largely reflect the way they are used in discourse (Coulmas 1981; Cowie 1988; Svensén 2009: 191). In some cases, dictionary equivalents can be provided, but in many cases, contrastive differences between the SL and TL cause untranslatability at word level. Consequently, lexicographers are forced to use co-text or context to show how an SL lexical item is reflected in the TL, which is also the case in the ESD.

In the entries for *gracious, great, holy, goodness* and *bless*, the hash is used to indicate the absence of equivalence of a pragmatic sense at word level. The examples illustrating these senses are idiomatic to a certain extent, which is also reflected in their idiomatic translations into Slovene. Given their idiomatic characteristics, they could be included in the idioms section of the respective entries. If that were the case, zero equivalence would not be an issue, since the sentential form has a perfect equivalent in the TL, i.e., in Slovene.

In many cases, several senses of a lemma in the SL are characterized by a lack of dictionary equivalence in the TL. The lexicographers compiling the ESD decided to include

all senses of the lemma in English with zero equivalence in Slovene under one sense, the only exception to this rule being the lemma *come*, where three senses have a hash sign. The reason for this exception is that two senses are characterized by a specific structure (sense 8, 'in questions after how'; sense 9, *come sth (with sb)*). The reduction in the number of senses with a hash seems a sensible decision and can be regarded as a way of simplifying the dictionary entry structure, which is a welcome feature in complex polysemous entries.

The contrastive analysis of the examples illustrating the semantics of the senses marked by the hash shows that it is not uncommon to find examples combining lexical and grammatical characteristics. This can be attributed to the distinct structure of languages resulting in the fact that they express different characteristics in different ways. A good example is the English verb *come* used in the patterns *come* + gerund and *come* + prepositional phrase (sense 12 in the ESD). The examples illustrating *come* in sense 12 are rendered into Slovene either by a perfective form of the verb (e.g., *prileteti* 'come flying') or by a structure which semantically expresses a completed action (e.g., *stopiti v veljavo, začeti veljati* 'come into effect'). It should be stressed that in Slovene, the grammatical aspect is expressed by the form of the verb: a verb may be perfective (e.g., *prileteti* 'come flying') or imperfective (e.g., *leteti* 'fly').

The lemmata *come* and *gonna* include a theoretical explanation providing grammatical information. In the ESD, theoretical explanations are short, precise, to the point, and above all, characterized by the use of very simple language. The metalanguage used in the ESD is Slovene, which also holds true for the short descriptions, since the dictionary is primarily intended for native speakers of Slovene. Short descriptions can be regarded as valuable, since they represent a short comment on the specific use dealt with in a specific sense.

5. Conclusion

In bilingual dictionaries, contrastive differences between the SL and TL, as well as features typical of either the SL or the TL, result in different types of equivalence. The focus of our study is on how bilingual lexicographers tackle the problem of zero equivalence at word level. One of the most important conclusions is that, if equivalence cannot be achieved at word level, but is possible at the level of the entire message, the problem can be resolved by including examples of use. Lexical items always occur with their collocates; consequently, one can expect them to appear in a similar co-text or context. Illustrative examples should be translated into the TL if the dictionary is to function as an effective communication tool. The user is not made aware of the problem of zero equivalence if the examples remain untranslated; consequently, a bilingual dictionary does not clearly show how two different languages function in everyday use.

References

Adamska-Sałaciak, Arleta. 2006. Translation of Dictionary Examples – Notoriously Unreliable? Corino, Elisa, Marello, Carla, Onesti, Cristina (eds.). 2006. *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6–9 settembre 2006*, Vol. 1. Alessandria: Edizioni dell'Orso. 493–501.

Adamska-Sałaciak, Arleta. 2010. Examining Equivalence. *International Journal of Lexicography* 23 (4). 387–409.

Al-Kasimi, Ali M. 1977. *Linguistics and Bilingual Dictionaries*. Leiden: E. J. Brill.

Atkins, Bertyl T. Sue. 1992/1993. Theoretical Lexicography and its Relation to Dictionary-Making. *Dictionaries* 14. 4–43.

Coulmas, Florian. 1981. *Conversational Routine: Explorations in Standardized Communication Situations and Prepatterned Speech*. The Hague: Mouton.

Cowie, Anthony Paul. 1988. Stable and Creative Aspects of Vocabulary Use. Carter, Ronald, McCarthy, Michael (eds.). 1988. *Vocabulary and Language Teaching*. London, New York: Routledge. 126–139.

Gouws, Rufus H. 2002. Equivalent Relations, Context and Cotext in Bilingual Dictionaries. *Hermes, Journal of Linguistics* 28. 195–209.

Jacobsen, Jane Rosenkilde, Manley, James and Pedersen, Viggo Hjørnager. 1991. Examples in the Bilingual Dictionary. Hausmann, Franz Josef et al. (Eds.). 1991. *Dictionaries. An International Encyclopedia of Lexicography*, Vol. 3. Berlin: de Gruyter. 2782–2789.

Svensén, Bo. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

Vrbinc, Alenka and Vrbinc, Marjeta. 2016. Illustrative Examples in a Bilingual Decoding Dictionary: An (Un)necessary Component?. *Lexikos* 26. 296–310.

Wiegand, Herbert Ernst. 2002. Equivalence in Bilingual Lexicography: Criticism and Suggestions. *Lexikos* 12. 239–255.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. Prague: Academia, The Hague, Paris: Mouton.

Zöfgen, Ekkehard. 1991. Bilingual Learner's Dictionaries. Hausmann, Franz Josef et al. (Eds.). 1991. *Dictionaries. An International Encyclopedia of Lexicography*, Vol. 3. Berlin: de Gruyter. 2888–2903.

Towards Building a Language Family Tree for Low-Resource Languages: Clustering Using Orthographic Features

Angelica Dela Cruz, Nathaniel Oco, Rachel Edita Roxas
National University – Manila, Philippines
delacruzah@yahoo.com

Abstract

In this paper, we present a novel approach, based on an ongoing study, to cluster 43 Philippine languages towards building an updated Philippine language family tree using orthographic features. The 43 languages are classified as follows: 17 are identified as developing, 13 as educational, 10 as wider communication, 2 as threatened and 1 as vigorous. Included in the 43 languages is Yami language, spoken in Taiwan but considered similar with Ivatan, a northern Philippine language. We used orthographic features our main source of data. In particular, we used character trigrams (3-gram), which are 3-character slices of a word. For example, the word "lexicon" will produce a trigram model of {"_le", "lex", "exi", "xic", "ico", "con", "on_"}. Our corpus consists of religious text from the holy Bible. We used existing applications to generate the trigrams per language and specifically used hierarchical clustering to build the hierarchy of languages based on feature similarity among languages. Our results are comparable to the language subgroups of Ethnologue. To further our analyses, a cognate list was also used to determine language similarity. As example, words like "hapon" (afternoon), "braso" (arm) and "dugo" (blood) were found in at least two languages from the same subgroup. The work can be extended by considering other techniques and features.

Keywords: Hierarchical clustering, trigrams, Philippine languages, cognate list

Introduction

According to Ethnologue[1], there are 187 listed languages in the Philippines, 41 are institutional, 72 are developing, 45 are vigorous, 14 are in trouble, 11 are dying and 4 are already extinct. The existing language family tree which can be found in Ethnologue were based from different studies in order to show information about Philippine languages. The study of Wurm [11] in 2007 was used as basis to determine if the specific language is endangered or not. The study of Crystal [1] in 2003 was used as basis of including English in the list of Philippine languages because it is used globally while the studies of Reid [10], Zorc [12], and Lobel [4], [5], [6] were used as basis for presenting the relationship between Philippine languages. These studies made serious efforts in constructing subgroup of Philippine languages with the use of different features such as morphology, phonologies, syntactic features and word list to represent the languages. But as time passes by, languages have a tendency to evolve and be influenced by its neighbor languages, especially its phonetic features, and might cause the need for Philippine language subgrouping to be updated. In addition, these studies were all conducted using manual means which are laborious and time consuming. Also, these studies are focused on specific language subgroups that are only part of the Philippine language family tree.

There are existing studies that conducted experiments to automatically measure similarity between languages. These includes [7], [8] and [9] but only limited languages were covered. Previous studies [2], [3] used various features such as phonetic, orthographic and geographical features to automatically cluster languages, however, the evaluation scores were still a bit low. In order to improve its results, the goal of this initial study as part of an ongoing study is to automatically cluster 43 Philippine languages towards building a Philippine language family tree using orthographic features. Character trigram (3-gram), which are 3-character slices of a word were used in this study to represent the orthography of the languages. For example, the word "lexicon" will produce a trigram model of {"_le", "lex", "exi", "xic", "ico", "con", "on_"}. A cognate list was also used to further the analyses.

[1] http://www.ethnologue.com/country/PH/languages

Methodology

A. Data collection

There are two orthographic features used in the study: trigram models and language word list. For the trigram models, online religious text documents of all the domain languages were collected. Considering the available resources gathered, the number of words used were only limited to 100,000, this is important to have fair results with all the languages gathered. The word list of the languages were given by the "Komisyon sa Wikang Filipino" that consists of 200-300 words for all the languages.

B. Data processing

The collected data were cleaned to remove all characters that are not necessary in generating trigrams such as numbers and punctuation marks. Regular expressions were utilized using Notepad++ in order to clean the data.

C. Clustering

The processed data were fed to a data mining tool to automatically cluster the languages using orthographic features. Hierarchical clustering algorithm was used in clustering the languages, it is a clustering method that builds hierarchy of groups of languages based on the similarity between the languages.

D. Evaluation

Resulting clusters will be evaluated using both extrinsic and intrinsic evaluation metric. As this is part of an ongoing study, the resulting clusters made from the orthographic features only are evaluated using an extrinsic evaluation metric first. For the extrinsic evaluation metric, we will use purity. This evaluation metric measures the validity of the clusters made by measuring the similarity of languages within a cluster based on an external expert knowledge. The ethnologue Philippine language subgrouping was used as the gold standard. Initial results are not yet evaluated using intrinsic evaluation.

Results and Discussion

Based on the initial results of the experiment, it can be observed that Yami and Ivatan are also found similar while Chavacano and Sama are considered as outlier languages. Our results, after evaluation, are comparable to the language subgroups of Ethnologue. To further our analyses, cognate list was also used to determine language similarity. As example, words like "hapon" (afternoon), "braso" (arm) and "dugo" (blood) in the Tagalog language are also present in the Cebuano language. Both Tagalog and Cebuano are under one language subgroup. The work can be extended by considering other techniques and features. Although the results of this study is already comparable to the language subgroup of ethnologue, there is still a need to collect data on other Philippine languages not yet covered by this study. Also, the results are still evaluated using external evaluation metric which uses external basis which may be outdated.

Acknowledgement

This work is supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (No. IIID-2015-07).

References

[1] Crystal, D. (2003). English as a global language.
[2] Dela Cruz, Angelica, Maria Cristina Co, Adrian Martin Sy, Nathaniel Oco. (2017). Building a Language Family Tree using Various Features. In Proceedings of the 17th Philippine Computing Science Congress.
[3] Dela Cruz, Angelica, Nathaniel Oco, Leif Romeritch Syliongka, Rachel Edita Roxas. (2016). Phoneme Inventory, Trigrams, and Geographic Location as Features for Clustering Different
[4] Lobel, J. W. (2004). Old Bikol-um-vs. mag-and the loss of a morphological paradigm. *Oceanic Linguistics*, *43*(2), 469-497.
[5] Lobel, J. W. (2005). The angry register of the Bikol languages of the Philippines. *Liao and Rubino*, 149-166.
[6] Lobel, J. (2013). Philippine and North Bornean Languages: Issues in Description, Subgrouping and Reconstruction.
[7] Oco, N., Syliongka, L.R., Roxas, R.E. (2016). Clustering Philippine Languages. In: 16th Philippine Computing Science Congress (PCSC).
[8] Oco, N., Sison-Buban, R., Syliongka, L.R., Roxas, R.E., Ilao, J. (2014). Trigram Ranking: Metric for Language Similarity and Clustering. Malay, pp. 53-68
[9] Oco, N., Ilao, J., Roxas, R.E., Syliongka, L.R. (2013). Measuring Language Similarity using Trigrams. 2013. International Conference on Recent Trends in Information Technology (ICRTIT).
[10] Reid, L. (1971). Philippine minor languages: word lists and phonologies. (Oceanic Linguistics Special Publication No. 8.) xiii, 241 pp. [Honolulu]: University of Hawaii Press.
[11] Wurm, S. A. (2007). Australasia and the Pacific. encyclopedia of the world's endangered languages, 425.
[12] Zorc, D. (1977). The Bisayan Dialects of the Philippines: Subgrouping and Reconstruction. Pacific Linguistics. Series C - No. 44. The Australian National University.

Framing specialized concepts through automatic extraction and semantic annotation: the DEFORESTATİON event

Beatriz Sanchez Cárdenas and Carlos Ramisch

LexiCon Research Group, Universidad de Granada (Spain) and

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille (France)

bsc@ugr.es and *carlos.ramisch@lis-lab.fr*

Abstract

Frame-based terminology is a linguistic framework whose main hypothesis is that specialized concepts and events are conceptually organized in the speaker's mind. It uses specialized semantic frames as a structured representation for this conceptual organization of domains. Lexicographic resources based on frame-based terminology represent a rich source of information that goes beyond static term inventories. Such resources can be useful for domain experts and translators, who need to have access to the domain's conceptual structures and corresponding linguistic realizations to be able to write idiomatic texts. Nevertheless, building lexical resources containing specialized frames can be time-consuming. Therefore, it requires the right tools and methodological framework to perform corpus queries, extract and analyze useful examples, and elicit the conceptual structure of a concept or event. This study presents a methodology designed to build up semantic frames for specialized concepts. The first step of our methodology is based on automatic corpus extraction of the linguistic participants involved in a concept. It is followed by a manual semantic annotation of verbs and noun phrases, thus allowing the inference of lexical structures that account for the argumental structure of the concept. We argue that, under certain conditions, integrating complex nominals in corpus extraction is essential to reach high-quality results, that will maximize their utility for the construction of the specialized frames. The DEFORESTATION event, in the domain of environmental sciences, is presented here as a case in point. We illustrate all steps of our methodology, including corpus queries, semantic annotation, until the creation of specialized frames for the DEFORESTATION event.

Keywords: frame semantics, phraseology, terminology, information extraction, semantic annotation

1. Background of the study

Frame semantics represents events according to the event's participants and their roles, reflecting its cognitive structure (Fillmore et al 2003, Fillmore 2006). Frame-based terminology (FBT) applies the premises of frame semantics to the study of the conceptual organization that underlies specialized domains (Faber 2012, 2015). According to this model, specialized semantic frames model non lexicalized senses across languages (Faber 2015). Frame-based lexicographic resources are a valuable resource, since they can be used to create specialized multilingual dictionaries, develop translation tools, and improve the translation of specialized texts.

Semantic frames provide a rich and structured framework to organize concepts in a specialized domain. Specialized semantic frames provide a way to cluster related lexical structures that account for language-independent dimensions of specialized knowledge. For example, a semantic frame representing the DEFORESTATION event will probably contain lexical units related to its consequences (e.g. *erosion, greenhouse effect, soil pollution*), as this is a relevant lexical structure environmental sciences.

The advantages of multilingual semantic frames for terminological and translation purposes are numerous. In fact, such a representation is a proxy that allows the inference of the cognitive structures underlying scientific texts. Getting to know the frame structures of these concepts and its linguistic correlates is useful both for translators, to understand a concept, and for non-native speaker experts, to produce idiomatic text in a foreign language.

2. Objectives

Given that frames are formed by complex argumental structures, their creation requires both linguistic and domain expertise, as well as tools for performing corpus-based searches (L'Homme et al 2014, Hermann et al 2014). Lexicographers building entries of frame-based resources need to have access to specialized corpora, so that they are able to perform complex searches, involving verbs and their arguments. Nevertheless, constructing lexical resources that model the frame structure of domain concepts is not only challenging but also a highly time-consuming task.

Despite some attempts, a systematic methodology for the creation of specialized semantic frames in specialized language is a challenging issue that has not been been solved yet, to the best of our knowledge. Ideally, computer tools could support, enhance and facilitate corpus analysis to confirm and generalize linguistic introspection. However, concordancers are not sufficient since one needs to run complex queries that are capable of modeling morphosyntactic and syntactic co-occurrence patterns that approximate predicate-argument structure. Moreover, variability can influence the results of corpus queries, so it is also important to take into account complex phenomena such as verbal alternation and complex nominals.

This article describes a protocol for the construction of specialized frames which encompasses the following steps: (i) design and application of complex corpus queries for triple extraction, (ii) systematic triple annotation by lexicographers, and (iii) semi-automatic triple grouping and manual frame construction. For the first step, triples composed of predicates and their arguments were automatically extracted from a specialized corpus with a view to creating language-independent specialized semantic frames. As a case study, we focus on the concept of DEFORESTATİON. The corpus used for this research was extracted from an English environmental corpus of 23 million words, and consists of a sub-corpus of 1,257,216 words composed of texts about the DEFORESTATİON event.

3. Hypothesis

These are the main theoretical assumptions underlying this study:

1. Frame-based terminology is a rich framework for representing specialized knowledge.
2. Specialized corpora contain information required for creating specialized frames.

Based on these two principles, we formulate the hypothesis that it is possible to automatically extract from corpora predicate-argument information that allows to generalize over different lexical realizations of predicates and argument classes, thus guiding the creation of language-independent specialized semantic frames.

From an environmental point of view, deforestation is one of the main ecological issues worldwide. From a terminological perspective, the concept DEFORESTATİON raises questions such as which entities and processes cause this event and what are their consequences upon the ecosystem. Some of the terminological resources dealing with this are EcoLexiCon[1], DiCoEnviro[2] and Gemet[3]. Although they all provide some useful information, none of them gives a thorough answer to these questions. Our approach tackles this problem from the perspective of frame-based terminology with the goal of enriching the mentioned resources.

4. Methodology

Our starting point is an English corpus of Environmental Sciences, containing more than 23 million words, including scientific articles, encyclopedic entries and specialized news. The present study focuses on a subcorpora containing 52,741 sentences and 1,257,216 dealing with concept and, thus containing the English term *deforestation*. The corpus was automatically preprocessed, including part-of-speech tagging, lemmatization, and dependency parsing using the UDPipe tool and the pre-trained models made available for the CoNLL-2017 shared task (Straka et al 2016, Straka and Straková 2017)[4].

In order to maximize the utility of the information extracted from the corpus, we have developed a dedicated tool for running corpus queries. This tool is based on the assumption that argumental structures relevant to a frame often occur in the prototypical form of "noun-verb-noun": a nominal phrase, followed by a verb or verb locution, followed by one or more nominal phrases. In the remainder of this section, we describe our methodology and the corresponding tool for automatic extraction (Section 4.1), then we detail the semantic annotation of the extracted elements (Section 4.2). The resulting DEFORESTATİON event frame structure is based on the semi-automatic grouping of annotated elements, described and analyzed in the next section.

4.1 Automatic extraction

Our methodology for automatic extraction makes use of MWEtoolkit: a computational tool originally designed to help creating general-purpose lexicons of multiword expressions (Ramisch 2015). The tool contains a powerful query engine that allows expressing multi-layer patterns such as *all sequences of nouns and adjectives preceding the word "forest"*. Therefore, it was easily adaptable to the context of specialized frame creation, since it enables

[1] http://ecolexicon.ugr.es
[2] http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi
[3] https://www.eionet.europa.eu/gemet/en/themes/
[4] http://hdl.handle.net/11234/1-1990

the extraction of phraseological units in texts using variable degrees of under-specification (i.e. with constraints on lemmas, parts of speech, and/or syntactic dependencies).

In this step, our goal was to extract relevant triples from the corpus. A *triple* is composed by a first nominal phrase (henceforth *n1*), followed by a verb or verbal locution (henceforth *v*), followed by a second nominal phrase (henceforth *n2*). For example, we expect to extract relevant triples such as [*deforestation, accelerate, global warming*] from our deforestation subcorpus.

A triple is a proxy that models the co-occurrence of predicates (*v*) with their corresponding arguments (*n1* and *n2*). Predicate-argument structures represented as simplified triples [*n1*, *v*, *n2*] will probably miss some useful information. For instance, some predicates are mostly nominal rather than verbal (e.g. *erosion* rather than *to erode*), and some predicates may have more or less than two arguments (e.g. [Soybean expansion] in southern Brazil [contributed] to [deforestation] by [stimulating migration to agricultural frontier regions]). Nonetheless, we expect that, by performing several queries using variants of the target terms, we can compensate for this inaccuracy of the triples and cover the full phraseological pattern that we aim to describe.

a) Complex nominals. One characteristic of the DEFORESTATİON event is that it involves participants which are lexicalized as complex noun compounds such as *global warming*, *forest loss*, and *greenhouse effect*. We will refer to these participants simply as *complex nominals,* to avoid discussing whether they are terms*.* A preliminary version of our tool allowed to locate triples where *n1* and *n2* were single nouns. However, we realized that it was not very useful to extract triples containing only the head nouns (e.g. *warming*, *loss*, and *effect*).

Therefore, as a preliminary extraction step, we run a query to extract recurrent noun phrases from the corpus. This query consisted on the following pattern, expressed as a regular expression over parts of speech:

(ADJ|(NOUN ADP?)){1,4} NOUN

As a result, it was retrieved from the corpora nouns (NOUN) preceded by a sequence of 1 to 4 modifiers ({1,4}), that can be either an adjective (ADJ) or another noun. In cases were the modifier is a noun, it can be optionally (?) followed by a preposition *(*ADP). This pattern extracted combinations such as *carbon dioxide*, *environmental harm*, *international environmental law*, and *victim of environmental harm*. Secondly, MWEToolkit was used to calculate the t-score association measure of each extracted combination (Evert 2004). The extracted combinations were ranked by descending t-score and those that seemed relevant to model the deforestation event were manually selected. The resulting list contains 94 recurrent complex nominals used in our further queries, and will be abbreviated as *list-complex* in what follows.

Furthermore, our corpus analysis showed that the DEFORESTATİON concept can be lexicalized by several term variants. Therefore, we manually elaborated a second list of terms corresponding to the usual denominations of this concept in English, such as *forest shrinkage*, *forest loss*, *forest scarcity*, *scarcity of forest*, and *land clearing.* This list contains 12 elements, and will be referred to as *list-deforestation* in what follows.

b) Triples extraction. Given the the two lists of recurrent complex nominals and variants of the deforestation term, the triples were retrieved by running queries in which at least one element of the triple was left under-specified (keyword *ANY*). MWEtoolkit was used

to extract from the corpus sequences "*n1-v-n2*" with at most three intervening words, as follows:

n1 []{0,3} v []{0,3} n2

The sequence []{0,3} is a placeholder for a sequence of three arbitrary words at most that can occur between the relevant elements, such as determiners or adverbs, and that will not be shown in the list of retrieved results. Elements *n1*, *v*, and *n2* can correspond to simple nouns and verbs, or to elements of *list-complex* or *list-deforestation*, as described above. For instance, a query such as *n1=list-complex*, *v=ANY*, and *n2=list-deforestation* will retrieve triples such as [*technological change*, *reduce*, *deforestation*] and [*population growth*, *associate*, *forest loss*]. In addition to simple verbs (e.g. *disrupt*, *accelerate*, and *cause*), we also extracted from the corpora a list of verbal locutions that often co-occur with terms related to deforestation (e.g. *lead to*, *result in*, and *depend on*). Finally, the queries resulting from the combination of all possible permutations (*list-deforestation* in position *n1* or *n2* and at least one under-specified element) was run. The extraction contained 598 triples, out of which 154 were considered as relevant for frame construction, which means a precision rate of 25,8%. The unselected triples suggest new research lines for future work. The selected triples were manually annotated as explained in Section 4.3. The error analysis hereunder explains the low percentage of relevant findings, thus allowing the enhancement of this protocol in future research.

4.2 Error analysis

The rate of errors in the triples retrieved is mostly due to the fact that DEFORESTATİON is a highly complex event that entails a large variety of interrelated participants. As shown below, some complex structures are difficult to identify automatically. These drawbacks will be addressed in future research.

1. Complex syntactic structures with three arguments such as an Agent (*deforestation*), a Patient (*acres of once production land*) and a Result (*desert*) where we obtained the misleading triple [*land*, *turn into*, *deforestation*]:
    - *Each year, **millions of acres of once productive land** are turned into **desert** through **overgrazing and deforestation.***
2. Deep semantic structures with a positive verb that hides a negation are problematic. For instance, the occurrence below retrieved the triple [deforestation, leave, tree], incorrect not only because of the negation but also due to the fact that this structure has three arguments:
    - *Deforestation leaves fewer trees to absorb carbon dioxide*.
3. Causal structures are difficult to identify since it requires to extract deep structure information. In the example below the triple [*activist, decry, deforestation*] misses the most relevant information. Future searches should contain knowledge-rich patterns, such as Sketch Grammars (León et al 2016).
    - *Environmental activists decried the apparent accelerating pace of deforestation in the twentieth century **because of** the potential loss of wildlife and plant habitat and the negative effects on biodiversity.*
4. Coordination of several nouns in one phrase. Currently only the first one of these nouns [e.g. *deforestation*] is detected by our MWEtoolkit scripts:
    - *The progressive conversion of the coastal land to alternative uses has been documented to cause **deforestation, pollution of marine and inland waters, coral reef destruction, coastal erosion and flood.***

5. Phrasal verbs are not yet correctly identified. This could be solved by detecting these verbal forms prior to running the searches:
   - *Cropper et al. (1999**) found** population pressure, road density and proximity to the capital city **as (found as=are)** the major factors responsible for deforestation in Thailand.*

4.3 Semantic annotation

As shown in Table 1, triples were annotated in three stages. Firstly, verbs were classified into lexical domains (Faber & Mairal 1999). Then, verb arguments were ranked according to their semantic class (e.g. LANDFORM, ACTİON, FLORA) according to a preliminary noun typology of environmental sciences, currently under development in our team. Lastly, verb arguments were assigned a thematic role from a closed inventory (e.g. Agent, Theme, Result). The goal of this three level annotation (verb lexical domains, nous semantic classes and noun thematic roles) is twofold. On the one hand, it reveals recurrent lexico-grammatical patterns in the corpora. For instance, the structure "ACTİON increase/stimulate PROCESS", where the first noun acts as Agent and the second as Patient is very frequent. Some of the nouns of the category ACTİON in this structure are *banana production, commercial ranching* and *soybean expansion,* whereas the PROCESS is lexicalized by *deforestation*, *land clearing* and *forest erosion*. On the other hand, semantic annotation of noun-verb-noun combinations allows the inference of recurrent conceptual schemes lexicalized that would be difficult to infer otherwise. In other words, this annotation decomposes specialized frame creation into systematic steps that are individually more tractable than the creation of the whole frame at once based on extracted triples. For instance, there are three main conceptual dimensions activated by this concept; (1) DEFORESTATİON begins to exist, (2) DEFORESTATİON is intensified and (3) DEFORESTATİON is seen as a direction towards which lead several processes and actions.

Based on the assumption that similar phraseological patterns reveal semantic similarity, the triples were automatically grouped into semantic classes using a program that groups triples with the same annotation. For instance, rows 3 and 4 of Table 1 will be grouped, as the verb's lexical domain, as well as the nouns' classes and roles, are identical. This information was then analyzed for the construction of the specialized frame as explained in 5.

NOUN1	N1 ROLE	N1 CLASS	VERB	LEXİCAL DOMAİN	NOUN2	N2 ROLE	N2 CLASS
demand for land	Patient	process>action	**spur**	CHANGE	**deforestation**	Agent	attribute
deforestation	Agent	process>change	**intensify**	CHANGE	**natural_flood**	Patient	process>loss
forest_clearing	Agent	process>loss	**contribute_to**	CHANGE	**change in biodiversity**	Patient	process>loss
forest_clearing	Agent	process>loss	**contribute to**	CHANGE	**climate_change**	Patient	process>loss
forest_degradation	Cause	process>change	**be a precursor of**	EXISTENCE	**deforestation**	Result	process>loss
forest_scarcity	Theme	attribute	**drive by**	MOVEMENT	**land_price**	Agent	process>action
forest_scarcity	Agent	attribute	**lead to**	MOVEMENT	**higher land price**	Result	process>action
technological_change	Cause	process>change	**promote**	EXISTENCE	**deforestation**	Result	process>loss
technological_change	Agent	process>change	**affect**	CHANGE	**forest_clearing**	Patient	process>loss

Table 1. Example of triples semantic annotation

5. Results

The annotated triples were grouped, based on their similarity and according to the three-layer annotation. Since all the verbs sharing the same kind of annotated lexical domain, thematic roles and semantic categories were put together, these lexical schemas reflected not only the lexico-grammatical patterns of *deforestation*, but also the conceptual structure of this concept. The grouped annotation indicates that there are at least three conceptual dimensions of DEFORESTATİON activated in scientific texts: EXISTENCE, CHANGE and MOVEMENT.

Table 2 shows the conceptual structure of the EXISTENCE lexical domain where the *deforestation* event is seen as the result of cause lexicalized by several actions (*banana plantation, labour, unsustainable development*), as the result of a change process (*maize boom price, loss of biodiversity, migration*) or as a theme that occur at certain places (*tropical area*).

LEXICAL DOMAIN	ARGUMENT 1			VERB	ARGUMENT 2		
	SYNTAX	THEMATIC ROLE	SEMANTIC CATEGORY		SYNTAX	THEMATIC ROLE	SEMANTIC CATEGORY
EXISTENCE	Subject	CAUSE	ARTIFICIAL PLACE	*associate to*	Direct Object	RESULT	PROCESS>ACTION
			ATTRIBUTE	*provoke, promote*			
			LANDFORM	*tend to, cause*			PROCESS>CHANGE
			METHOD	*produce*			
			PROCESS>ACTION	*cause*			PROCESS>LOSS
			PROCESS>CHANGE	*provoke, promote, cause*			
		RESULT	ATTRIBUTE	*be a precursor of, promote, result in, encourage, determine*	Direct Object	CAUSE	PROCESS>ACTION
			FLORA	*result_from, result from*	Passive Subject		
			PROCESS>ACTION	*promote*	Direct Object		PROCESS>LOSS
			PROCESS>CHANGE	*result_from*	Passive Subject		
			PROCESS>LOSS	*depend_on, promote, explain_by, encourage, result, generate*	Direct Object/Passive Subject		
		THEME	LANDFORM	*result from*	Circumstantial	LOCATION	PROCESS>CHANGE
			PROCESS>CHANGE	*occur_in*			PROCESS>LOSS
			PROCESS>CHANGE	*occur_in*		THEME	
			PROCESS>LOSS	*correlated with*		LOCATION	
			PROCESS>LOSS	*occur_within*		THEME	

Table 2. Lexical domain of EXISTENCE in DEFORESTATION

Interestingly enough, although one of the main Environmental problems worldwide is to prevent and revert deforestation, the scientific texts of our corpora do not deal with the issue of what should be done for protecting forests and ensuring sustainable forest management. Further research should corroborate whether this is due to a biased corpora or if the scientific community is not sufficiently addressing this issue.

Finally, future studies, should take into account interlinguisitic correspondences in order to know how this frame is lexicalized in other languages.

6. Discussion and conclusions

The lexical schema obtained from the analysis of the corpus triples extraction reflects the conceptual structure of DEFORESTATİON. We have showed how to extract from a specialized corpus the information that lexicalize its conceptual frame Our study shows that semantic frame of the DEFORESTATİON event entails a complex scenario in which many participants are interdependent. This semantic frame represents a process caused by a human

activity or a natural event. The results indicate that human activities are accountable for this process either directly (by cutting trees) or indirectly (because of greenhouse gas emissions causing natural disasters that result in soil erosion). The consequences of deforestation affect the whole ecosystem at many levels and have a significant impact on living organisms.

Our study has also uncovered many interesting linguistic structures that are currently missed by our automatic extraction procedures. Besides, it would be necessary to replicate this procedure in other languages, in order to stablish interlinguistic correspondences. These will be a topic for future research, with the final goal to give access to more relevant conceptual as well as linguistic information in terminological resources.

7. Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness, and was partly funded by project PARSEME-FR (ANR-14-CERA-0001), and by the PARSEME Cost Action (IC1207).

8. References

Cognitive linguistics: Basic readings, 34, 373-400. *International Journal of Lexicography*, 16(3), 235-250. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1448-1458). *LREC.* 2014.*Theory and Applications of Natural Language Processing series*, XIV. Springer. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.*Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

Evert Stefan. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis. Institut für maschinelle Sprachverbeitung, University of Stuttgart, Germany.

Faber Pamela, and Ricardo Mairal. 1999. *Constructing a Lexicon of English Verbs*, New York, Mouton de Gruyter.

Faber, Pamela (eds). 2012. *A cognitive linguistics view of terminology and specialized language* (Vol. 20). Walter de Gruyter.

Faber, Pamela. 2015. Frames as a framework for terminology. *Handbook of Terminology*, 1(14). ed. by Kockaert, H.J. & Steurs, F., 1:14-33. John Benjamins Publishing Company.

Fillmore, Charles J. 2006. "Frame semantics". *Cognitive linguistics: Basic readings*, 34, 373-400.

Fillmore, Charles, Christopher, Johnson, and Miriam, Petruck. 2003. "Background to FrameNet". *International Journal of Lexicography*, 16(3), 235-250.

Hermann, K. M., Das, D., Weston, J., & Ganchev, K. (2014). Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1448-1458).

L'Homme, Marie-Claude, Benoît Robichaud, and Carlos Subirats Rüggeberg. "Discovering frames in specialized domains." *LREC.* 2014.

León Araúz, Pilar, Antonio San Martín and Pamela Faber. 2016. Pattern-based Word Sketches for the Extraction of Semantic Relations. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, 73-82. Osaka, Japan: COLING 2016.

Ramisch, Carlos. 2015. "Multiword Expressions Acquisition: A Generic and Open Framework". In *Theory and Applications of Natural Language Processing series*, XIV. Springer.

Straka, Milan, and Jana Straková. 2017. "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe". In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

Straka, Milan, Jan Hajič, and Jana Straková. 2016. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Translingual Words and Social Media: Are they Korean Words or English Words?

Brittany Khedun-Burgoine
Dr Jieun Kiaer
In consultation with Dr Danica Salazar
University of Oxford/Oxford English Dictionary
brittany.khedun-burgoine@orinst.ox.ac.uk

Abstract

In the age of global interconnectivity aided by the proliferation of digital technologies such as smartphones and social media, the transcultural flow of language is quicker than ever before. Consumers, previously limited to consuming local media in their native language, can now access a wealth of foreign language media instantaneously in the digital age.

Besides bringing a new level of global cultural awareness, social media and the internet have had a linguistic impact on language and the identity of words – we can now observe an increase in *translingual* words (Kiaer forthcoming), words which are not bound by being attributed to their language of origin. These translingual words now exist in multiple different languages and even in different varieties of World Englishes. As these words have crossed beyond their native language we now have to think of ways to define them in the additional foreign languages which they exist in.

We aim to conduct a content analysis of comments on social media platforms such as Twitter and to assess how these words have been mediated through fandom as *fandom lexicon* before settling in World Englishes. Using dictionaries such as the Oxford English Dictionary and other regional dictionaries we will examine how these words are currently documented, and the relationship between translingual words and the dictionary.

Keywords: Fandom Lexicon, Translingual Words, Social Media

Introduction: Social media and the globalisation of Korean-born words

In July 2012, Korean singer Psy released his 18th single "Gangnam Style", featuring the catchy lyrics of *oppan Gangnam style* repeatedly throughout the song and chorus. In early September 2012 it had overtaken Girls' Generation's "Gee" as the most viewed K-Pop video on YouTube, and by the end of the month it had topped the iTunes charts in 31 countries. This was a song in the *Korean* language, gaining prominence in English-speaking countries and topping the charts over incredibly popular English language songs released at the same time such as Carly Rae Jepsen's "Call Me Maybe". There was no doubt that this would not have been possible without the assistance of digital technologies and social media. After all, how else would a Korean pop song become a number one single in the United Kingdom?

With the music video's iconic "invisible horse dance" shared and parodied by popular celebrities such as Britney Spears, the viral success of "Gangnam Style" wasn't just spreading a fun, catchy song and music video – it was also spreading the Korean relational term *oppa*, "older brother" far beyond Korean borders (Lee 2015). Worldwide searches on Google for "oppa meaning Korean" began to peak at a higher frequency than ever before, with search popularity in September 2012 nearly seven times as much than in May 2012. The word *oppa*, a Korean kinship term meaning "older brother", features a total of eleven times in "Gangnam Style", and this prevalence has suddenly brought an unprecedented global interest in the *meaning* of it. Five years later in 2017, the search frequency for *oppa* reached the highest levels yet – showing a direct correlation with the explosive popularity of boy group BTS in the West. Korean-origin vocabulary continues to gain prominence.

English definitions of *oppa* have existed online long before the word initially peaked in search frequency after "Gangnam Style" went viral. *Oppa*, a Korean-born word, had already begun its journey into English, mediated through the English-speaking international fans of Korean pop music – known as K-Pop. *Oppa* has now become a *translingual* word (Kiaer forthcoming) that cannot be defined purely by its linguistic origin in Korean, as it now exists in English too. Much like people, translingual words are those which cross borders between languages and challenge the idea of a native, nation state lexicon. We are now moving away from a monolingual static lexicon model to a multilingual, ***dynamic*** lexicon model as the question of translingual words and how best to define them in the digital age becomes increasingly more important.

Objective

Most major digital technologies and social media platforms appeared in the mid 2000s between 2004-2007. Twitter, Instagram, Facebook and even the iPhone all appeared during this time period. This study aims to show that social media has aided the transcultural flow of information and language in the digital era, and this has not been without impact on the individual linguistic repertoire of the user. Consumers are no longer limited to local media and are able to access foreign media instantaneously, increasing their exposure to foreign-born words.

As globalisation continues to promote this sharing of lexicon and dispersion of foreign-born words, it begins to have an impact on their linguistic identity. Traditional notions of "borrowing" and "loanwords" are now increasingly less efficient ways of describing foreign-born words which have become prominent in other languages. In the case of translingual words which have origins in East Asian languages, there are often layers of cultural specific meaning, and to class them as "borrowed" or "loaned" would not be an

adequate way to describe words which have undergone vast semantic change on their journey into English (Kiaer forthcoming).[1]

With the ever-increasing popularity of Korean pop culture products and the "Korean Wave", [2] Korean-born translingual words like *oppa* are on the rise, and we seek to argue that they are the future of our lexicon, as they question the notion of a "native" language lexicon versus a more diverse and multilingual lexicon. Translingual words are a valuable asset to enriching lexical variety, and we aim to examine ways in which Korean-born translingual words are currently defined in the dictionary and understood on social media. This study will provide a valuable insight into how Korean-born words are mediated through speakers of World Englishes on social media, which impacts the semantic change and variation of definition between the native Korean and English.

Methodology

The methodology for this study will primarily consist of an analysis of the usage and definitions of a selection of four translingual words which are commonplace on social media, used most frequently by English-speaking members of the international K-Pop fandom. The four words are as follows:

Oppa, Girl Crush, Chingu, Skinship

These words have been selected as they provide a mixture of Korean-born vocabulary (*oppa*, *chingu*), Hybird English words (*Skinship*) and English-origin words which have become popular in Korea (*Girl Crush*). All of the selected words have numerous definitions on websites especially dedicated to the understanding of fandom lexicon and have featured in both *Urban Dictionary* and *The K-pop Dictionary*, cementing their status as important items of fandom lexicon.

OED methodology

In order to track and assess the usage of these words outside the online sphere of the international fandom, evidence will be gathered through online databases such as ProQuest, JSTOR and Google Books. Using these databased will provide evidence of widespread use which predates the fandom usage of the selected words prevalent on social media. Nexis, an online database of global digital newspapers, will also be used to search for the selected words as keywords, allowing for cross-referencing. Besides providing evidence of the first recorded instance of the words being used in English, this methodology can also help determine where these words fit into different parts of English speech and whether there have been any orthographic changes.

[1] It is worth noting that most loanwords in English are referring to objects, not relational or other terminology with specific cultural connotations barring a few exceptions. The OED contains the Japanese relational term *sensei*

[2] The Korean Wave refers to the increase in global popularity of Korean pop culture since the 1990s. The Korean Wave includes Korean pop music (K-Pop), Korean dramas (K-Dramas), Korean food (K-Food) and Korean cosmetics (K-Beauty)

Social Media

As these words are a distinct feature of online-based lexicon, part of the methodology for this study will consist of content analysis of comments featuring the selected words made on popular social media platforms, with a particular emphasis Twitter. As Twitter allows to search by hashtag and features accurate time stamping for each Tweet, we will be able to track any potential linguistic developments over the past ten years. As relatively less data is available pre-2008, we will be using data from the period of 2008-present[3].

Google Trends and the GloWbE

The popularity of Korean media is not limited to the "inner circle" speakers (Kachru 1986) of English, and as such Google Trends and the Corpus of Web-Based Global English (GloWbE) will be used to provide location-based popularity data for the usage of the selected words. Due to the growing influence and popularity of the Korean Wave in Southeast Asia, English-speaking fans in those regions are *instrumental* in the spread of Korean-origin vocabulary into World Englishes. Making use of Google Trends and the GloWbE can help assess these usages in varieties of English found in the "outer circle" of English speakers, as they become increasingly more important in the diversification of English lexicon.

Findings

Oppa: Older brother or a hot guy?

The Korean language, steeped in respect language to convey a complex system of social hierarchy, has a number of versatile kinship terms which can be used in and out of a familial context to convey closeness and intimacy, which English lacks. The kinship term *oppa* "older brother" can be a term of address for a blood relative, a close friend or even a romantic partner. Perhaps the most important aspect of the native Korean usage of *oppa* is that it is an age and gender-sensitive word. *Oppa* is used exclusively by a *younger female* towards an *older male*. It is this erosion of the age-sensitive factor in the English usage of *oppa* which will be focussed on in this study.

Following the OED methodology, the earliest usage of *oppa* in English comes from academic writing on a generative study of discourse in Korean and English written in 1972. The text notes that the usage of *oppa* as "female's elder brother" in lieu of a second-person pronoun (which would be the case in English) is *systematic* and *obligatory* in the Korean language. In an article published in 1983 by *The Associated Press*, instead of opting to translate *oppa* as "older brother", it is simply rendered in the article as "oppa (brother)". Interestingly, in an article published by *The Korea Herald* in 1998, the writer notes that a woman is talking to her *boyfriend* and yet again "oppa (brother)" is written. Despite the clear inclusion of the age-sensitive information in early academic writing, in the earliest appearance of *oppa* in English language media the age-sensitive information is missing – as is an attempt at translating *oppa*.

The situation on social media following the popularity of the Korean Wave however, shows a complete eradication of the age-sensitive factor – and in some cases, even the gender-sensitive aspect. Searching for *#oppa* on Twitter brings up more versatile and dynamic usages of oppa:

[3] Twitter is open to the general public for academic purposes and no identifiable information has been included in Tweets featured in this study

"Girl where was an #oppa when I needed one? HAHAHA! I always said I'd stay true to the fiery Latin American telenovelas with its scruffy men with washboard abs, but damn girl that viewpoint got OPPA-cified"
"starting my day being an #oppa"

It is fairly clear that *oppa* in these cases does not necessarily have any type of age-sensitive factor. Additionally, *oppa* is also referred to as *an oppa* – suggesting that *oppa* is now one singular type of person, bringing us further away from the several different types of person that an *oppa* could be in the native Korean. This type of usage is most prevalent in Philippine English, where the popularity of Korean pop culture media has led to *oppa* being associated with a handsome or attractive man, the age-sensitive information irrelevant. This is supplemented with data from Google Trends, where worldwide search frequency for "oppa meaning Korean" from the time period of 2004-present was most concentrated in the Philippines.

Chingu and Chingus

Despite the age-sensitive aspect of *chingu* not having the same importance as it does in *oppa*, *chingu* "friend" is used primarily for somebody of a similar age. You could for example, cause offence if you called somebody much older than you *chingu*, instead of using a more appropriate respect term. *Chingu* makes its first appearance in an English language newspaper in an article published by the *Korea Herald* in 1999. The author, an Indian man, writes of his blossoming friendship with a young Korean man. He is told "sir, you chingu. I discount", along with being affectionately touched on the hand. Early use, therefore, is not too dissimilar from the native Korean.

On social media, *chingu* has now become a prominent feature of fandom lexicon, especially in *fan to fan* interaction. Fans will frequently address each other as *chingu* in lieu of the English "friend". Yet again, much like in the case of *oppa*, the age-sensitive aspect which is so critical in Korean is removed in this fandom usage of *chingu*. Whilst the usage of *chingu* in lieu of "friend" is without translation error, it is often used without any other prior friendly interaction, which is most common on Twitter. Additionally, *chingu* is often also used alongside "friend" rather than simply replacing it.

The most peculiar difference in the usage of *chingu* in English is that fans tend to pluralise *chingu* in line with English orthography. Whilst it is possible to pluralise in Korean with the addition of the suffix *deul*, writing *chingu deul* in romanisation when you want to address multiple people is long winded. As a result of this, many fans instead choose to use *chingus*, adding "s" at the end to symbolise pluralisation, providing a simpler, more concise alternative.

Some examples below taken from Twitter under the *#chingus* hashtag:

"This is the best thing on Earth! I love my #chingus"
"I love u my non-judgemental friends. #chingus"
"my chngz!! #chingus"

What is most notable about the use of *chingu* or *chingus* is that it is an example of *core borrowing*. English already has a perfectly good word in "friend" which is interchangeable with *chingu*. One possible reason for this use is that using Korean-origin vocabulary identifies fans as part the fandom *in-group*, effectively showing that they have a command of the origin language of K-Pop and provides them with *fandom cultural capital* (Hills 2002).

Girl Crush

"Girl Crush" is a phrase which developed in the early 2000s in the English language and data from the GloWbE shows that out of a total 105 hits for "Girl Crush", 77 hits were from native English-speaking countries. Being a slang term, one of the earliest usages is from 2003 in an article in the *New York Observer*. A girl crush is defined as a "nonsexual mating ritual" and that it is the first step towards friendship with a fellow woman.

The word began to enter Korean fairly recently and is now largely credited with the girl group MAMAMOO. Their 2016 single "You're the Best", the lead single from their album *Melting* was released to critical acclaim and cemented their position as a popular girl group. On the same album featured a track called "Girl Crush", with the lyrics placing an emphasis on being *cool*, *awesome* and *sexy* as opposed to innocent or cute[4]. Google Trends data shows that searches for *"걸크러쉬 girl crush"* in South Korea peaked in February 2016 – coinciding with the release of *Melting.*

These lyrics about seeming cool, independent and sexy as opposed to cute and innocent are representative of the semantic change the phrase has undergone. Girl crush in fandom lexicon has now become an adjective, used to describe a certain type of girl, or a certain type of concept. Certain idols have become associated with the girl crush aesthetic, and certain groups, particularly MAMAMOO, are frequently regarded as a "girl crush group". Much like how in the original English definition of girl crush, the girl crush concept is designed to appeal to female fans as opposed to male.

The girl crush concept has become commonplace in a number of girl groups and has become synonymous with groups with a larger number of female fans than male. In English usage, we can now observe these two different versions of girl crush existing synonymously, as the Korean adjectival use of "girl crush" has entered back into the English language having been mediated through fandom lexicon. Popular girl groups MAMAMOO and Sistar performed at a concert dubbed as a "girl crush concert". Girl crush is now on its way to becoming a "boomerang word" – a word which has been altered in a foreign language and then returned back to its origin language. In this case, girl crush originated in English slang usage, before undergoing semantic change in Korean adding an adjectival usage and then returning to English through fandom lexicon as an *adjective.*

Skinship: From mother child maternal intimacy in Japanese to platonic intimacy in Korean to same-sex intimacy in English

Skinship has had an interesting journey being mediated through Japanese and English before settling into English. The exact origin of skinship is unclear – but earliest sources suggested that it originated in Japanese as *wasei eigo*[5] – a Japanese coined English word. Having originated in Japan before migrating to Korea then to the international fandom through the Korean Wave, skinship has now become a mainstay of fandom lexicon used on social media. Skinship is a blended word formed of the noun "skin" and the suffix "-ship", bringing to thought words expressing some level of personal intimacy such as "relationship", "friendship" and "kinship". The etymological origin of skinship expresses that skinship is the "physical intimacy of the skin".

[4] MAMAMOO "Girl Crush" lyrics https://colorcodedlyrics.com/2015/09/mamamoo-mamamu-girl-crush (Accessed 2018)

[5] 和製英語 wasei-eigo literally translated "Japanese-made English" are Japanese language expressions based on English

Skinship appears in English language material for the first time in a 1974 article by Donald L. Smith about "Ingrish" (Japanese English), where it is simply defined as the "physical contact between mother and child". A further article published by The Associated Press in 1982 makes reference to the "skinship of public bathing". The first time skinship crops up in a Korean context is in a Korea Herald article published in 1999, where it is described as "the Korean affinity for touching one another". From these definitions, we can see that skinship is a form of physical intimacy, with an emphasis on *skin to skin* contact. The early definitions in a Japanese context suggest special reference to communal bathing and physical intimacy between mother and child, but in the first definition in a Korean context we see the addition of this idea that in Korean culture there exists an "affinity for touching one another", and this is what constitutes *skinship*.

Skinship can be used in and out of a platonic context to suggest a very specific type of physical intimacy – two women or two men, for example, could perform acts of skinship together, as could a boyfriend and girlfriend. It is this romantic connotation which has taken prevalence in the fandom interpretation of the word. Whereas this usage is purely platonic, in the fandom lexicon usage of skinship it is most frequently used in reference to same-sex physical intimacy which might cross the boundary into *romantic* interest.

Examples collected from Twitter under the *#skinship* hashtag illustrates this usage:

*"chen's **skinship** is too gentle (again, baek let chen did it) wonder what were you guys talking to? this is too romantic #baekchen"*
*"Was Jinyoung's hand creeping in Mark's pocket for warmth, or also for **skinship**? #markjin"*

Both *#baekchen* and *#markjin* are combinations of two male idols names in what is often termed a "shipping name". Shipping is when fans support the relationship either platonic or romantic of two idols, and in the case of K-Pop fans many of these so-called "ships" are same-sex, as K-Pop is very gender segregated to avoid media attention.

Conclusion

As we can see from the examples above, these words challenge the idea of a "native" English language lexicon and serve to effectively cross nation state borders which exist between languages. English is an incredibly diverse language and as the lingua franca, many English words have gained localised meanings across the world and in World Englishes – we cannot judge the forms and meanings of the words used across languages from one single branch of English speakers. Even as the lingua franca, it does not mean that everybody will speak the same variety of English.

This is evident in the different fandom lexicon usages of the words included in this study. Fans, especially those in Southeast Asia, have played an integral and important role in mediating the translingual journeys of these words. This plays into Eckert's theory of the *third wave of sociolinguistic study*, which places the speakers in an active role of constantly enhancing and stylistically tailoring their linguistic repertoire – whereas previous sociolinguistic study placed greater emphasis on sociocultural factors influencing one's linguistic repertoire rather than greater emphasis on the speaker's *individualised* and *dynamic* lexicon. Fans may share their interest in Korean pop culture, but they do not share the same languages and they might not necessarily speak the same variety of English.

The relative grammatical freedom of language on social media and the internet furthermore offers English-speaking fans to not simply *receive* these words, but to also become major participants in the discussion of how to render Korean-origin words

appropriately in English, both orthographically and semantically. This process is *interactive* and *dynamic* – bringing us into a new era where English is moved out of a monolingual "native speaker" model and embraces multilingualism. This is crucial in the way we seek to define these translingual words in future. There is now a more complex and nuanced linguistic identity to translingual words – and it is not difficult to see how in our incredibly globalised, multilingual and diverse world – these translingual words are the future of our lexicon.

Bibliography

Abrams, J. (1982, September 16). TODAY'S FOCUS: Communal Baths Being Sunk By High Costs, Affluence. *The Associated Press*. Retrieved May 20, 2018 from www.lexisnexis.com/uk/nexis

Chang, S. (1972). *A Generative Study Of Discourse With Special Reference To Korean And English* (Order No. 7309899). Available from ProQuest Dissertations & Theses Global. (302618179). Retrieved from https://search.proquest.com/docview/302618179?accountid=13042

Cho, S. (1998, August 03). Seoul Searcher; Hand Phone Craze. *The Korea Herald*. Retrieved May 20, 2018, from www.lexisnexis.com/uk/nexis

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology, 41*(1), 87-100. doi:10.1146/annurev-anthro-092611-145828

Hills, M. (2002). *Fan cultures*. London: Routledge, Taylor & Francis Group.

Hwang, K. (1983, July 19). TODAY'S TOPIC: Popular Korean Television Show Reunites Families. *The Associated Press*. Retrieved May 20, 2018, from www.lexisnexis.com/uk/nexis

Kachru, B. B. (1986). The power and politics of English. *World Englishes, 5*(2-3), 121-140. doi:10.1111/j.1467-971x.1986.tb00720.x

Kiaer, J. (forthcoming). *Translingual Words: An East Asian Lexical Encounter With English.* London: Routledge, Taylor & Francis Group.

Lee, S., & Nornes, M. (2015). *Hallyu 2.0: The Korean wave in the age of social media*. Ann Arbor: University of Michigan Press.

Marshall, A. (2003, September 15). The Girl Crush: Manhattan Minxes Fall Under Spell. *New York Observer*. Retrieved May 20, 2018, from www.lexisnexis.com/uk/nexis

Smith, D. L. (1974). Ribbing Ingrish: Innovative Borrowing in Japanese. American Speech, 49(3/4), 185. doi:10.2307/3087797

Sullivan, J. (1999, August 10). 'Maximum Korea' offers author's personal observations. The Korea Herald. Retrieved May 20, 2018 from www.lexisnexis.com/uk/nexis

Tandon, A. (1999, August 28). My Korean friend. *The Korea Herald*. Retrieved May 20, 2018, from www.lexisnexis.com/uk/nexis

Some Observations on Collocation Dictionary of Adjectives in Turkish*

Bülent ÖZKAN
Mersin University
ozkanbulent@gmail.com

Abstract

In the present study, the collocational structures of adjectives, which are lexicological units in Turkish, are investigated. In this perspective, the data sets of the "Collocation Dictionary of Adjectives in Turkish" (CDAT) were evaluated on the basis of corpus linguistics data of lexicology. The mentioned corpus consists of 25 million words and the adjectives of Turkish were filtered from this corpus. Secondly, selected adjectives were compared with the 'Turkish Dictionary (TD), because the data set of TD is formed with traditional methods. As a result, the findings below are obtained about the adjectives of Turkish:

- Some adjectives that are defined as headwords in TD, are not found in corpus query,
- Some headwords that are defined as adjectives in TD, need to be added new meaning entries,
- Some headwords that are tagged as adjectives in TD, need to be tagged with extra lexical information.

Key Words: Adjectives, Turkish Language, Corpus Linguistics, Lexicology

* This study is based upon a National Research Project, supported by TÜBİTAK-SOBAG. Project number is 109K104 and titled as "Collocations of Adjectives in Turkey Turkish -A Corpus Based Application-". We appreciate the contributions of TÜBİTAK.

1.Introduction

Today the principles and methods of corpus linguistics make significant contributions to the studies of lexicology, which makes researches on the process of forming dictionaries and updating in time through the findings. Today, computational linguistics, also known as *Natural Language Processing* (NLP), by taking the language models called as corpus in parallel to applied linguistics, is commonly used in the studies of lexicology, grammar, dialect, science of translation, historical grammar and linguistic alternation, language teaching and learning, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and poetics (McEnery et al., 2006: 80-122; Kennedy, 1998: 208-310).

It is observed that parallel of the development of information technologies, the linguistic studies become enlarged. Today, existing methods and applications provide linguistic researchers with favorable opportunities via the computers on linguistics studies. Computers can offer linguistic researchers wider opportunities to provide broader linguistic data for various purposes.

Corpus linguistics is used extensively in all areas of linguistics studies, and this approach is a method that increases its validity day by day about description of any language, processing real-time data and modeling of language fit for the research questions. As known, compilation linguistics researches are very effective to reach experimental results in on language studies.

Especially in the study of lexicology, this situation is clearly observed. When the dictionaries that are created by using traditional method (*categorizations, personal preferences, rewriting etc.*) compared with the dictionaries that are created by using corpus linguistics, it can be seen that the positive contributions to the field of lexicology can be clearly observed.

In this perspective, this study is aimed to compare two different dictionary creating approaches: one of them is traditional dictionary of Turkish (TD) and other one is Collocation Dictionary of Adjectives in Turkish (CDAT).

2. Purpose of the Study

The aim of the study is to present the general corpus profile of **12.320** adjectives defined as headword in TD. Data set queried from a corpus of 25 million (+/-) words, formed through the principles and methods of corpus linguistics.

3. Population-Sample

As a population of corpus with 25 million (+/-) words, gathered from various thematic texts, which belong to the literary language of Turkish, and from the Internet environment by using varied software, is used. The corpus consists of **A- Printed Works** (%60) between the years of 1923-2008 and **B- Internet Texts** (%40) between the years of 2006-2008 (See also Table 1).

Table 1. *Content of Corpus*

A- PRINTED WORKS					
	LAYERS	**VARIANCE**		**%**	
1	Novel	96		22,802	
2	Poem	68	7*	16,152	
3	Tale	49	4*	11,638	
4	Essay-Critics	44	3*	10,451	
5	Theatre	35	1*	8,313	
6	Memoir	21		4,988	
7	Research	20		4,750	**%60**
8	Conversation-Interview-Article	18	1*	4,275	
9	Humor	14		3,325	
10	Travel Writing vb.	10	1*	2,375	
11	Letter	4	1*	0,950	
12	Biography	4		0,950	
13	Diary	1		0,237	
14	Various Types	30		7,125	
		403	**18***		
	TOTAL		**421**	**100**	
**Anthological works*					

B- INTERNET TEXTS				
	LAYERS	**SUB LAYERS**	**%**	
1	News etc.	*Politics, Economy-Finance, World-Live, Weather, Forecast, Sports…*	60	
2	Life	*Technology, Education, Tabloid Press…*	10	**%40**
3	Culture-Art-Health	*Health, Book, Cinema, Theatre…*	10	
4	Essay	*Column…*	20	
			100	
				100

4. Research Questions

According to the corpus queries of adjectives, the research questions are;

- According to corpus query what are the possible reasons of the not found headwords that are defined as adjectives in TD,
- Which headwords that are defined as adjectives in TD, need to be added new meaning entries and what are the possible reason of this.
- Which headwords that are tagged as adjectives in TD, need to be tagged with extra lexical information and what are the possible reason of this.

5. Findings and Interpretation*

When the headwords that are defined as adjective in TD and CDAT were compared:

a. Some adjectives defined as headwords not found in corpus query:

The numbers of **3.093** headwords that are defined as adjective in TD, are not found in corpus query. Among the possible reasons for the case of not being found, although the content of the corpus is a matter of the fact, as in the previous studies on defined lexeme in vocabulary (Özkan, 2010) the reasons such as the lexeme's being old and for this reason their being archaic, lexeme's belonging to a specific field, being a lexeme from slang or folk speech, wrong part of speech tagging can be listed (Özkan, 2014). Same reasons are valid for headwords that are defined as adjectives in TD.*

Sample of old/archaic headwords: abat, aharlı, alayişli, âlimane, alüfte, amelî, ayyar, barudi, berhayat, bihuş, bikarar, bivefa etc.

Sample of specific field headwords: aerolojik, akromatik, androsefal, andemik, avurtlu, aposteriori, astropikal, biyometeorolojik etc.

Sample of slang or folk speech headwords: acırak, afal, ağzı gevşek, andavallı, angın, apaz, apışak, apışık, başı devletli, bet suratlı, cambul cumbul, çirişçi çanağı, çiroz, çorlu etc.

Sample of wrong part of speech tagging: abone, asortik, basketçi, bloke, dekore, dibek, dolayı, fırın, hamur etc.

b. Some headwords defined as adjectives in TD, need new meaning entries:

With the result of the corpus query, new entries are added to the meaning of the adjectives. The number of 637 headwords' meaning reorganized and also added new meaning according to usage of adjectives.

According to corpus query, it's seen that, in some headwords that are defined as adjectives, the meaning of adjectives needs new meaning entries as metaphorical usage (See also Figure 1. and 2.).

Sample of adjectives: *acemi, acı tatlı, açık, acımtırak, ağdalı, ağır, ağır yaralı, ağırlıklı, ağızdan dolma, ağrılı, ağrısız, akışkan, aksak, aldatıcı, alt, altın, amatör, bayat, bayıltıcı, belalı, belgesiz, benden, beş altı, beşiz, besleme, betimleyici, bezeli, bilinçlendirici, bilinçsiz, bir karış, bir küme, bir nebze, birincil, bitişik, bitkisel, bizden, bloklu, bölmeli, boş, boyalı, boyasız, bozuk, bronz, bukağılı, bulandırıcı, bulaşık, bunaltıcı, buruk, büyük, büyük boy, büyükçe, çekici, çelik, çeşnili, çevresel, çeyrek, etc.*

Figure 1. Sample of "bayat" (stale).

* Because of eight (8) pages limit, adjectives have been given with limited sample. See also project report.

* For full list of not found adjectives from Özkan, B. (2011). TÜBİTAK-SOBAG-109K104 "Collocations of Adjectives in Turkey Turkish -A Corpus Based Application-" Project Report. (http://uvt.ulakbim.gov.tr/uvt/index.php?cwid=12&vtadi=TPRJ&vt_no=0&j_no=714&year=2012)

Figure 2. Sample of "kırık" (broken).

c. Some adjectives need extra lexicological information:

It's observed that some headwords, which are defined as adjective in TD, took place necessarily with some certain morphemic components. That lexemes which are defined as adjective like: *...adlı, ...bacaklı, (-ile) bağlantılı; ...sayfalık, ...yıllık, ...aylık* are lexicalized with certain morphemic structures. On the other hand, the lexemes like *-de/-ile yetişmiş, ...kaplama, (-i/-ile) dolmuş* are lexicalized with verbal structures and also some lexemes which are constructed with several colour and numbers morphemic components (Özkan, 2018). Number of 380 adjectives need extra lexicological information like below:

(i) Used with any genitival structure: ...başlıklı, ...yapraklı, ...görünümlü, ...adlı, ...ruhlu, ...beyazlı, ...giyimli, ...burunlu, ...yüzlü, ...elbiseli, ...saçlı, ...dolusu, ...bandıralı, ...kafalı, ...desenli, ...kalpli etc.

(ii) Used with only case suffixes: sarılı (~ile, ~e, ~de), çevrili (~ile), kalan (~den, ~e), uygun (~e), ilgili (~ile), ilişkin (~e, ~ile), benzer (~e), kumandalı (~den), haiz (~e, ~i), dair (~e), kalma (~den), yetişmiş (~de, ~den), takılı (~e), örtülü (~ile, ~i), gelen (~e, ~den) etc.

(iii) *Used with a number:* şişe, dönümlük, yıllık, buçuk, tarihli, bardak, sayılı, milyarlık, senelik, asırlık, aylık, paket, puanlı, sayfalık, saatlik, liralık, litre, beygirlik, sekiz etc.

(iv) Used with case suffixes and any genitival structure: evvelki (~den, ~…), endeksli (~e, ~…), örülü (~ile, ~…), bezeli (~ile, ~…), dolu (~ile, ~…), asılı (~e, ~de, ~…), tahsisli (~…, ~e), zıt (~e, ~…), sonraki (~den, ~…), kaplı (~ile, ~…), önceki (~den ~…), tutulmuş (~e, ~…), perdeli (~i, ~ile, ~…), dolmuş (~i, ~ ile, ~…), ayrı (~…, ~den), eğilimli (~…, ~e) etc.

(v) *Used with a number and any genitival structures:* odalı (number, ~...), saniyelik (number, ~...), ayaklı (number, ~...), sermayeli (number, ~...), yataklı (number, ~...), bacaklı (number, ~...), başlı (number, ~...), numaralı (number, ~...), kişilik (number, ~…), çocuklu (number, ~...), basamaklı (number, ~...) etc. and

(vi) Used with -mAsI morphemic structure: -mAsI olanaksız, -mAsI güç etc.

6. Conclusion

Every single language go through changes in terms of vocabulary. Naturally, in every language, new words are derived and come into use consistently. In this study, lexemes, specified as adjective in TD, are investigated through an extensive literary language corpus in terms of *corpus query, meaning entry and extra lexicological information.*

As a result of corpus query of adjectives, **12.320** adjectives analyzed in this study, it is seen that the number of **3.093** adjectives are not found as "in usage" in TD. It has possible reasons, one of the reasons of this is lexicalized headwords from archaic/old vocabulary, second reason of this is their belonging to a specific filed and the other one is being from slang or folk speech, the last one is wrong part of speech tagging.

Result of the corpus query, new entries are added to the meaning of the adjectives. The number of **637** headwords' meanings are reorganized and also added new meanings according to usage of adjectives.

On the other hand, number of **380** adjectives need extra lexicological information because of their view of usages. These are classified as ***(i)*** *used with any genitival structure* ***(ii)*** *used with only case suffixes,* ***(iii)*** *used with a number* ***(iv)*** *used with case suffixes and any genitival structure* ***(v)*** *used with a number and any genitival structures* ***(vi)*** *used with -mAsI morphemic structure.* Also these morphemic structures should be tagged for a user-friendly dictionary.

The study is significant in terms of presenting the methodological content for the regeneration of a usage-based dictionary of TD in line with the principles of corpus linguistics and lexicology.

Acknowledgement

This study is based upon a National Research Project, supported by TÜBİTAK-SOBAG. Project number is 109K104 and titled as "Collocations of Adjectives in Turkey Turkish -A Corpus Based Application-". We appreciate the contributions of TÜBİTAK.

References

Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics.* New York: Addison Wesley Longman Limited.

McEnery, Tony et al. (2006). *Corpus-Based Language Studies an Advanced Resource Book.* New York: Routledge.

Özkan, B. (2010). Güncel Türkçe Sözlükte Zarf Olarak Tanımlı Sözlükbirimlerin Derlem Denetimi. *Turkish Studies International Periodical for the Languages, Literature and History of Turkish or Turkic.* Volume 5/3 Summer 2010: 1764-1782.

Özkan, B. (2010). *Turkish Corpus,* Mersin University.

Özkan, B. (2011). TÜBİTAK-SOBAG-109K104 "Collocations of Adjectives in Turkey Turkish -A Corpus Based Application-" Project Report. (http://uvt.ulakbim.gov.tr/uvt/index.php?cwid=12&vtadi=TPRJ&vt_no=0&j_no=714&year=2012)

Özkan, B. (2014). "Türkiye Türkçesi Söz Varlığında Fiillerin Derlem Denetimi ve Derlem Tabanlı Sözlüğü." *bilig. Türk Dünyası Sosyal Bilimler Dergisi*. Sayı: 69. Spring 2014. 1719-204.

Özkan, B. (2018). Adjectives as Lexicological Components at Teaching of Turkish Language as Mother and Foreign Language Education and Language. *International Journal of Education and Language*. Winter 2016 1 (1): 1-6.

Turkish Dictionary: http://tdk.gov.tr/

Türkçe Sözlük (2005). Ankara: TDK Yay.

http://tdk.gov.tr/

http://turkcederlem.mersin.edu.tr

A Multi-dimensional Discrimination of the English Equivalents in Chinese-English Dictionaries for Chinese Users

Chengmin Liao, Lixin Xia, Mengyu Zhang
Guangdong University of Foreign Studies
983135093@qq.com CDHUIYI@ALIYUN.COM

Abstract

This paper conducted an investigation into the definitions in 9 popular Chinese-English dictionaries for Chinese users, and it is found that they generally discriminate the English equivalents by means of illustrative examples, collocations, pragmatic labels, usage notes and cross references. However, on close inspection, some common problems are detected among these discrimination methods. Firstly, the means to discriminate the equivalents are monotonous. The existing Chinese-English dictionaries mainly rely on illustrative examples to discriminate them. But common users can hardly tell the subtle differences in the synonymous words of the target language simply by one or two examples. Secondly, the objects of discrimination are not clear. Some Chinese-English dictionaries set up special discrimination columns, but their objects of discrimination are synonymous Chinese headwords rather than the English equivalents which dictionary users need most. With regard to Chinese users, these discrimination columns are of little help in their English production. Thirdly, the manner of discrimination is not explicit. The existing Chinese-English dictionaries generally provide limited discrimination information in the collocations, usage notes, pragmatic labels and cross references, but they do not discriminate the meaning and usage of the English equivalents in a more explicit way, such as establishing discrimination columns, error warnings, and explanatory notes, etc. Finally, the paper proposes the basic principles for the equivalent discrimination: to cater for the needs of dictionary users comprehensively, provide multi-dimensional discrimination information explicitly, and discriminate the English equivalents rather than the Chinese headwords. Sample entries are given to illustrate the methods and strategies for the equivalent discrimination.

Keywords: Chinese-English Dictionary, Equivalent, Discrimination, Bilingual Lexicography

1. Introduction

Many Chinese users of Chinese-English dictionaries have a common feeling that the existing Chinese-English dictionaries which, in most cases, provide only the English equivalents of Chinese headwords, haven't paid enough attention to discrimination of the English equivalents. Consequently, dictionary users always need to look up the dictionary again and again for a single headword in that they can hardly determiner to choose which one when faced with a group of English equivalents sharing similar meaning. For this purpose, this paper will investigate common methods on discrimination of English equivalents in existing Chinese-English dictionaries for Chinese users, analyze their deficiencies and propose suggestions for improvement.

2. Current situations on discrimination of the English equivalents in existing Chinese-English dictionaries

This paper selects 9 Chinese-English dictionaries for Chinese users as research objects, of which selection criteria are dictionary types, the extent of influence and usage rate among dictionaries users. First of all, with their great influence in China and large scope in each revision, the first edition of *A Chinese-English Dictionary* (Wu Jingrong, 1980) (hereinafter referred as *C-E 1st Edition*), published after reform and opening up, as well as *A Chinese-English Dictionary (Revised Edition)* (Wei Dongya, 1995) (hereinafter referred as *C-E 2nd Edition*) and *A Chinese-English Dictionary (Third Edition)* (Yao Xiaoping, 2010) (hereinafter referred as *C-E 3rd Edition*) are selected as objects of our research. Then, as for small size and large size dictionaries, two portable Chinese-English dictionaries, *Concise English-Chinese Chinese-English Dictionary* (Feng Juehua, 2001) (hereinafter referred as *Concise C-E*) and *A Modern Pocket Chinese-English Dictionary* (Chen Hongan, 1990) (hereinafter referred as *Pocket C-E*), and another famous for large amount of word collection, *The Chinese-English Dictionary* (Wu Guanghua, 2010) (hereinafter referred as *Large C-E*), are chosen respectively. *A Practical Chinese-English Dictionary for Translation*, compiled by Wu Wenzhi and Qian Housheng in 2001 (hereinafter referred as *C-E Translation*), is taken as representative of translation dictionary. In addition, *A New Century Chinese-English Dictionary* (Hui Yu, 2004) (hereinafter referred as *New Century*) is highly praised with its focus on dictionary practicability and central role of readers, providing especially much more collocations than other Chinese-English dictionaries. *New Age Chinese-English Dictionary* (Wu Jingrong & Cheng Zhenqiu, 2005) (hereinafter referred as *New Age*) is welcome by lexicographers and reader with its principle of seeking innovation, accuracy and practicality. Therefore, this paper considers these 9 Chinese-English dictionaries as research objects.

Existing Chinese-English dictionaries generally discriminate the subtle differences between English equivalents by means of grammatical notes, collocations, usage notes, special columns, cross references and error warnings, as in Table 1 below.

Table 1: Main manners of discriminating the English equivalents in Chinese-English dictionaries.

Number	Dictionary Name	Illustrative Example	Grammatical Note	Collocation	Usage Note	Pragmatic Label	Cross Reference	Discrimination Column	Error Warning
1	C-E 1st Edition	+	-	+	+	+	+	-	-
2	C-E 2nd Edition	+	-	+	+	+	+	-	-
3	C-E 3rd Edition	+	-	+	+	+	+	-	-
4	Large C-E	+	-	+	-	+	+	-	-
5	Concise C-E	+	-	+	-	+	+	-	-
6	Pocket C-E	-	-	-	-	-	-	-	-
7	C-E Translation	+	-	+	-	+	+	-	-
8	New Century	+	-	+	+	+	+	-	-
9	New Age	+	-	+	+	+	+	-	-

By inspecting the existing Chinese-English dictionaries, we have following findings:

Firstly, illustrative example is the most frequently used manner in discriminating English equivalents of each Chinese-English dictionary. Among our 9 dictionaries, all other dictionaries, except pocket-sized *Pocket C-E*, supply illustrative examples to explain specific meaning and usage of English equivalents in particular contexts. Especially when there are several English equivalents sharing close meanings, provided with appropriate illustrative examples, dictionary users would be able to infer those subtle differences in meaning and usage among English equivalents.

E.g. 1 事件shìjiàn 名incident; event: 流血~ *blood incident*/意外~*accident*/二十世纪最大的政治~之一*one of the greatest political events of the 20th century*

from *C-E 3rd Edition*

In example 1, *C-E 3rd Edition* offers two English equivalents of the headword "事件" and three illustrative examples. Perhaps, some careful users may infer from these illustrative examples that equivalent *incident* is inclined to denote negative things while equivalent *event* to denote significant things in historical and personal life, which, as a matter of fact, is the main difference of meaning between these two near-synonyms. They both could denote various kinds of things, but in shades of meaning, *incident* is often used with negative words, such as *serious*, *violent*, *unfortunate*, *pollution*, *unpleasant*, etc. On the contrary, positive words such as *historical*, *major*, *social*, *important*, *significant*, etc. are usually employed by *event*.

Secondly, except *Pocket C-E*, all other dictionaries provide a few collocations, which are generally given within round brackets "( )". The co-occurring subjects, objects and other types of collocations are usually annotated in English in some dictionaries, such as *C-E 1st Edition*, *C-E 2nd Edition*, *C-E 3rd Edition*, *C-E Translation*, *New Century* and *New Age*.

E.g. 2 【磕磕绊绊】kēke-bànbàn <形>①(of a road) bumpy; rough:…. ②(of a person) limping; jerky:…

from *New Century*

In example 2, *New Century* indicates the co-occurring subjects, in English, of the English equivalents by glosses.

Large C-E and *Concise C-E* give glosses in Chinese before English equivalents, which are as follows:

E.g. 3 上课[- -] ①（学生听课）attend class; go to class…②(教师讲课)conduct a class; give a lesson…

from *Large C-E*

In example 3, the usage contexts of English equivalents are annotated in Chinese in Large C-E, the method of which is almost similar to that employed in Concise C-E. However, the information annotated in these two dictionaries doesn't seem to serve for collocations. To indicate the collocations, "学生" and "教师" would be enough, with "听课" in "学生听课" and "讲课" in "教师讲课" being redundancies. As a result, these annotations may be the Chinese definitions of the Chinese headword. On a close inspection of *Large C-E*, we could find that it is a special characteristic of this dictionary to adopt both Chinese definition and English equivalent, of which, in almost each single word entry, the Chinese definition is first given, and then the English equivalent. Moreover, in multi-word entries, Chinese definition is selectively given for certain senses, and then English equivalents, such as example 4 below. Sometimes, collocation is implied in Chinese definition, as example 3 above, nevertheless, in most cases, there is no collocation implies. Therefore, based on these phenomena, probably we could conclude that *Large C-E* doesn't annotate collocation on purpose, but only contains some elements of collocation occasionally in Chinese definition.

E.g. 4 平常[-] ①（普通；不特别）ordinary; common…②(平时)generally; usually; ordinarily; as a rule…

from *Large C-E*

On researching the collocations of dictionaries mentioned above, we also find that collocations offered in existing Chinese-English dictionaries don't play obvious function in meaning discrimination since they have generally divided equivalents with semantic distinctions into separate senses. For example, in example 2 and 3, there are two senses respectively, of which the collocations are annotated before each entry after the entry number, indicating that both or all equivalents under that entry are governed by the collocation. In fact, in example 2, *bumpy* and *rough* under the first entry as well as *limping* and *jerky* in sense 2 have distinct collocates. *Bumpy* collocates most frequently with *ride*, while *rough* with *ground*; *limping* collocates most with *person*, while *jerky* with *movement*. Dictionary users would be misled by such ambiguous collocations in example 2, so the existing Chinese-English dictionaries need to improve further their collocation from the perspective of discrimination on equivalents.

Thirdly, all dictionaries, except *Pocket C-E*, employ pragmatic labels and cross references for discrimination of English equivalents. Subject labels are most frequently used among pragmatic labels, for instance, 【药】(Medicine),【哲】(Philosophy),【电】(Electricity),etc. Some dictionaries also give stylistic and register labels like [褒](Commendatory), [口](Spoken), [文](Written), etc. All these labels play an important role in discriminating English equivalents. However, the form and content of cross references in each dictionary are relatively simple, most of which provide the cross reference by simply listing the corresponding Chinese headword, for example in C-E 3rd Edition, under the entry of "牝", the cross reference of its antonym is indicated in the form of "(opp. 牡)".

Fourthly, five dictionaries including *C-E Dictionary (1-3 Edition)*, *New Century* and *New Age* all adopt usage notes, which are provided as glosses before or after the English equivalent. See following examples:

E.g. 5 【捉摸】zhuō mō<动>[usu. used in the negative]fathom; …

from *New Century*

E.g. 6 捉摸zhuōmō动fathom; ascertain[usu. in the negative]…

from *C-E 3rd Edition*

E.g. 7 事宜shìyí (usu. used in official documents, laws and documents, etc.) matters concerned; relevant matters…

from *New Age*

From these instances we could see that usage notes on existing dictionaries are used to explain the usages of Chinese headwords in that these indicated usages don't apply to their corresponding English equivalents. For example, English equivalents *fathom* and *ascertain* are not labeled with "usually used in the negative" in each English dictionary, while "捉摸" is explicitly labeled with "（多用于否定句）(often used in the negative)" in *Modern Chinese Dictionary*. Therefore, we could conclude that usage notes on existing Chinese-English dictionaries are aimed at annotating the source language, that is, Chinese. On the contrary, what Chinese students need most, the usage notes of the English equivalents, are not offered at all in these dictionaries, consequently providing limited help in discrimination of the English equivalents.

Fifthly, other three methods, grammatical notes, discrimination columns and error warnings, which are most frequently used to express meaning in English Learner's Dictionaries, are not employed in our 9 dictionaries. This may demonstrate that existing Chinese-English dictionaries still take providing English equivalents, but not discrimination of English equivalents, as their principal compiling purpose and think dictionary users themselves could search for the grammatical information and meaning discrimination of the English equivalents in English-Chinese dictionaries and monolingual English dictionaries. Nevertheless, some dictionaries compilers have recognized the importance of discrimination of English equivalents, for instance Guanghua Wu sets special discrimination columns in some other Chinese-English dictionaries compiled by him, like *The New Chinese-English Dictionary* and *A Comprehensive Chinese-English Dictionary*.

However, on a close analysis of discrimination columns in *The New Chinese-English Dictionary* and *A Comprehensive Chinese-English Dictionary*, we find that the objects of discrimination are semantic differences of Chinese headwords, for instance:

E.g. 8 对照[-zhào]（对比；参照）contrast; compare; cross-reference; collation…

【辨析】**对照**[duìzhào]，**对比**[duìbǐ]：它们都有"相互比较"的意思，"对照"侧重"使二者的矛盾性质更加鲜明突出"，此外，还有"互相对比参照"的意思，…"对比"通过比较，"突出事物之间的区别、差距"，…

from *A Comprehensive Chinese-English Dictionary*

In example 8, we could notice that the discrimination column provided here only makes semantic difference between Chinese headwords "对照" and "对比". The compiler probably aims at providing information of meaning discrimination for foreign Chinese learners since this dictionary is intended to serve for both Chinese users and foreign users.

However, Chinese users need to know, besides the English equivalents of "对照", the differences of meaning and usage among *contrast*, *compare*, and *cross-reference*. Discrimination of Chinese headwords "对照" and "对比" will hardly benefit Chinese dictionary users in English production, whereas foreign users might need this kind of information. Anyway, the major users of this type of so-called dictionary "both for Chinese and foreign users" are eventually Chinese people. As a result, intending to cover both native and foreign users in one dictionary may finally come to nothing at all.

In the end, dictionary trying to pay attention to all aspects must be very huge. As for *A Comprehensive Chinese-English Dictionary*, it is divided into 3 volumes with more than

6,000 pages altogether. It will always take pains to look up a word as dictionary users need to judge which volume this word belongs to, which is inconvenient to common dictionary users. In short, most Chinese-English dictionaries discriminate the English equivalents by means of illustrative examples, collocations, pragmatic labels, usage notes and cross references, among which illustrative examples are most widely used. These methods do help dictionary users in discriminating English equivalents to some extent, but the existing Chinese-English dictionaries mainly rely on illustrative examples to discriminate English equivalents while common users can hardly tell the subtle differences in the synonymous words simply by one or two examples. More explicit manners, like discrimination columns, error warnings, etc. of discrimination, if employed, will contribute to users' better understanding on discrimination of English equivalents. Nevertheless, no existing Chinese-English dictionary adopts this explicit manner. Some Chinese-English dictionaries, such as *The New Chinese-English Dictionary* and *A Comprehensive Chinese-English Dictionary*, set up special discrimination columns, but their objects of discrimination are synonymous Chinese headwords rather than the English equivalents which dictionary users need most. With regard to Chinese users, these discrimination columns are of little help in their English production.

3. Suggestions on discrimination of the English equivalents in Chinese-English dictionaries

Chinese-English dictionary compilers often list more than one English equivalent which cannot be regarded as fully equivalent to the headword. Perhaps the combination of these equivalents is what the headword means or each equivalent reflects a part of headword's meaning, which will bring communication barriers to learners. Chinese-English dictionary compilers can't be satisfied with these partial equivalents, but should multi-dimensionally reveal the subtle differences on meaning and usage of the English equivalents so as to eliminate communication difficulties for dictionary users.

First, the objects of discrimination in Chinese-English dictionaries for Chinese users should be the English equivalents rather than their Chinese headwords. Dictionary users possess relatively limited knowledge of the target language, English, and can hardly tell those subtle differences between or among English equivalents. On the contrary, Chinese headwords, in their native language, Chinese, need not to provide further explanation.

Then, Chinese-English dictionary should try its best to put discrimination information in a more explicit manner. There is no doubt that illustrative example is a commendable form to show usage contexts. However, in most cases readers could not effectively select the appropriate English equivalent by mere examples. What dictionary users want is that the dictionary is capable of explaining the differences of equivalents in a concise and explicit way for their convenience.

Third, the English equivalents in Chinese-English dictionaries are usually a group of near-synonyms which are close or related in meaning, but they may differ in referential meaning, associative meaning, grammatical meaning, collocative meaning and pragmatic meaning. Therefore, more methods should be adopted to discriminate the equivalents, including semantic discrimination, grammatical notes, pragmatic labels, collocation, discrimination columns and error warnings, etc.

E.g. 9 板bǎn <名>**board；plank**（厚木板）**；plate**（金属板）**；sheet**（金属薄板）：3厘米厚的木板a board three centimeters thick.…

In example 9, *board* is the prototype equivalent of the Chinese headword "板", while other equivalents (*plank*, *plate*, *sheet*) are its hyponym concepts with different referents.

Under this circumstance, annotating distinctive features of these hyponyms within glosses can help discriminate equivalents clearly, which is widely used in Chinese-English dictionaries.

E.g. 10 事件shì jiàn<名>**e'vent**（常指具有重大意义的大事）；**'incident**（常指具有消极影响的事件）；**'accident**（尤指意外发生的事故）：家中发生的意外事件accidents in the home||二十世纪最大的政治事件之一one of the greatest political events of the 20th century||这显然是一起非常不幸的事件。It is obviously a very unfortunate incident.…

In example 10, the three English equivalents share same referential meanings but diverse associative meanings and affective meanings. More specifically, *event* possesses positive semantic association, often collocating with words such as *historical*, *important*, which have positive semantic prosodies. *Incident* tends to have negative association, often collocating with *serious*, *unfortunate*, *unpleasant*, which have negative semantic prosodies, while *accident* denotes something unexpected. Annotating the associative meanings of these equivalents in glosses may assist learners in comprehending subtle semantic differences among the English equivalents, which is convenient for using and memorizing these vocabularies.

E.g. 11 擅长 <动>**be good** (at sth./ at doing sth.)**; ex'cel** (**-ll-**) *vi.* (in/at sth./ at doing sth.)：小女孩擅长钢琴。The little girl is good at playing the piano.|| 这个队擅长打防守反击。The team excels at turning defence into attack.…

In example 11, the two English equivalents share same referential meanings and associative meanings but diverse grammatical meanings. First of all, the prepositions following them are different: *be good* could only be followed by *at*, while excel by either *in* or *at*. Secondly, when expressing "擅长做某事", we could only adopt *excel at doing sth.* but not **excel in doing sth.* The above-mentioned grammatical collocations could not only discriminate the English equivalents, but also help learners to better master the grammatical properties of these two words.

With regard to the entry "磕磕绊绊" from *New Century* cited in the preceding part of this paper (see e.g. 2), we believe that although the entry offers collocations of the English equivalents, such as "(of a road) bumpy; rough" in sense 1 and "(of a person) limping; jerky" in sense 2, this manner of discrimination is still not perfect enough. Dictionary users may be misled that *bumpy* and *rough* in sense 1 share identical collocation without any distinction, and so do *limping* and *jerky*.

However, by searching Sketch Engine, we find words co-occurring with *bumpy* are as follows: *ride* (22), *road* (11) and *track* (8). And words co-occurring with *rough* are *ground* (60), *ride* (53), *track* (40), *sea* (34) and *grass* (22). Furthermore, the frequency of *bumpy* in corpus is 156 while *rough* is 3291. Therefore we could conclude that bumpy is low-frequency, mainly describing unevenness of land route, while rough mainly describes unevenness of soil, as well as that of sea, grassland, etc.

In addition, for *jerky* and *limping*, their collocations emphasize different aspects. Jerky, collocating with *movement* (23) and *motion* (2), primarily describes unsteadiness of action, but *limping*, collocating with people (2) and walk (2), principally describes people's walking. As to their frequency, *jerky* is 90 and *limping* is 35. Therefore, according to above research, we should modify the entry as follows:

E.g. 12 磕磕绊绊kēke-bànbàn <形>①(of a ground, ride, grass,etc.) **rough**; (of a ride, road, track, etc.) **bumpy**:…. ②(of a movement, etc.) **jerky**; (of a person, walk, etc.) **limping**:…

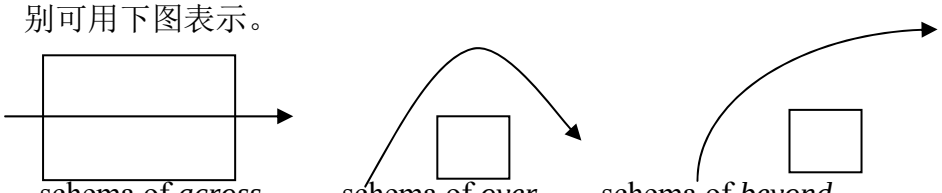
In example 12, we provide collocations of two equivalents under the same sense respectively. Careful users are supposed to find their differences on collocation and usage. On the other hand, we adjust the sequence of the English equivalents based on frequency, with high-frequency one first. Lastly, collocates are also ranked from high to low frequency, and "etc." is added to show collocations provided in the entry are just some typical ones so dictionary users can still choose other appropriate collocations in accordance with context. In a word, Chinese-English dictionaries for learners should pay attention to more detailed collocations in order to discriminate the differences of equivalents in collocative meanings.

E.g. 13 律师lǜshī〈名〉**lawyer**; **attorney**/əˈtəːni，əˈtɜːni/<美>; **counsel**【律】: 被告/原告的律师今天出席了庭审。The counsel for the denfence/procecuting was present in court today.…

In example 13, the three English equivalents share same referential meanings but diverse pragmatic meanings. Among them, *lawyer* is lingua franca, which could be used in both British and American English. *Attorney* could only be used in American English, especially applied to job titles, which is more formal than *lawyer* in American English. And as for *counsel*, it is a legal term, denoting the lawyer who represents a party in court. By explaining pragmatic differences with register label "<美> (Ame.)" and subject label "【律】(Law)", dictionary users are able to correctly choose the right word to produce English, and they could better comprehend the differences and similarities between British and American English.

E.g. 14 越过yuèguò〈介〉**across**; **over**; **beyond**: 越过公园go across the park||越过地平线go beyond the horizon||狗越过篱笆。The dog jumped over the fence…

辨析：across、over和beyond都有"越过，穿越"的意思，across表示越过某一平面，有障碍物不能自由穿越时不能使用。over是越过某一障碍物，越过某一平面时不能使用。beyond有越过某一障碍物，到"很远的地方"之意，如My house is just beyond the river.与My house is just across the river.相比，前者有在河的那一边很远的意思。它们的区别可用下图表示。



schema of *across* schema of *over* schema of *beyond*

In example 14, the three equivalents share same grammatical functions but diverse connotative meanings. First, *across* is used with verbs, indicating that someone or something departs from one side of a plane, traversing the plane, to the other side. The area traversed must be a plane. If it is a line, *cross* cannot be used, and neither do a plane blocked by barrier. Then, *over* is also used with verbs, generally indicating that someone or something departs upwards from one side of a barrier, getting over the barrier after reaching the peak, then fall downwards to the other side of the barrier. However, if we want to express "汽车驶过大桥", we can say both "The car drove across the bridge." and "The car drove over the bridge." with the former emphasizing that the bridge is a plane which the car traverses while the latter focusing on the river under the bridge and that the car, passing above the river, reached the other side of the river. Last, *beyond* indicates that someone or something moves upwards from one side of an object and reaches the other side, but continuing to extend forward very far. In this example, discrimination column is set to discriminate these English equivalents and image schemas of each equivalent are presented respectively. This visualized method may help dictionary users better understand those subtle differences among English equivalents, stimulate their interests and strengthen their memories.

4. Conclusion

Chinese-English dictionaries for Chinese users are active dictionaries with English learners and English teachers whose native language is Chinese as target users. From the perspective of users, this type of dictionary should principally provide information about the target language, English. Nevertheless, in the aspect of discrimination of the English equivalents, the existing Chinese-English dictionaries give much more attention to the Chinese headwords, thus leading to low usage rate of the provided information among users. Certainly, the above-mentioned deficiencies of Chinese-English dictionaries are summarized from the view of their target users. However, from the view of dictionary typology, the biggest problem with compilation of Chinese-English dictionaries in China is compilation idea. The existing Chinese-English dictionaries in China still try to meet all needs of various kinds of dictionary users within one dictionary, emerging so-called dictionary both for Chinese and foreign users. As a result, dictionaries are becoming increasingly bigger and thicker, which is not accepted by dictionary users yet. For these reasons, we should alter our traditional compilation ideas and compile diverse types of Chinese-English dictionaries specific to different dictionary users, for instance, Chinese-English dictionary for translation specific to users engaged in English translation, and Chinese-English dictionary for learners specific to English learners.

In this paper, basing on actual needs of dictionary users, we propose that Chinese-English dictionaries for learners should take the English equivalents as focuses of discrimination and utilize various methods, such as illustrative examples, glosses, labels, columns, illustrations and usage notes, to discriminate the subtle differences of the English equivalents in referential meaning, associative meaning, grammatical meaning, collocative meaning and pragmatic meaning. In our sample entries, we use only one or two discrimination methods to explain how to carry out discrimination of the English equivalents, but in practical lexicography, we could adopt diversified methods according to practical situation so as to better demonstrate the differences between or among the English equivalents.

5. Acknowledgement

This paper was funded by the "Innovation Project of Guangdong University of Foreign Studies for Training International Postgraduate Talents".

6. References

Atkins, B. T. S., & Varantola, K. (1997). Monitoring Dictionary Use. International Journal of Lexicography, (1): 1-45.

Chen Guohua & Tian Bing. (2008). Design Features of the Next Generation of English Learner's Dictionaries. Foreign Language Teaching and Research, (3): 224-233.

Chen Hongan. (1990). A Modern Pocket Chinese-English Dictionary. Beijing: Modern Press.

Feng Juehua. (2001). Concise English-Chinese Chinese-English Dictionary. Changchun: Jilin University Press.

Gao Houkun. (1988). The *Chinese-English Dictionary* of 1978 Viewed in Retrospect. Foreign Language Teaching and Research, (2): 65-68.

Hui Yu. (2004). A New Century Chinese-English Dictionary. Beijing: Foreign Language Teaching and Research Press.

Rundell, M. (1998). Recent Trends in English Pedagogical Lexicography. International Journal of Lexicography, (4):315-342.

Rundell, M. (2009). The Latest Development and Future of Corpus Lexicography (□&□) (Xia Lixin & Zhu Dongsheng, Trans.). (Original Paper Published in 2008). Lexicographical Studies, (3/4): 71-78/81-91.

Su Xin. 2004. New Compilation Ideas on the Fourth Generation of Chinese-English Dictionary—A Review on *A New Century Chinese-English Dictionary*. Foreign Language Education, (2): 6-8.

Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Wei Dongya. (1995). A Chinese-English Dictionary (Revised Edition). Beijing: Foreign Language Teaching and Research Press.

Wu Gunghua. (2010). The Chinese-English Dictionary (Third Edition). Shanghai: Shanghai Translation Publishing House.

Wu Jingrong. (1980). A Chinese-English Dictionary. Shanghai: The Commercial Press.

Wu Jingrong & Cheng Zhenqiu. (2005). New Age Chinese-English Dictionary. Shanghai: The Commercial Press.

Wu Wenzhi & Qian Housheng. (2001). A Practical Chinese-English Dictionary for Translation. Guilin: Lijiang Publishing House.

Xia Lixin. (2011). Reflections on Compilation and Publication of Chinese-English Dictionaries in China. Publishing Journal, (2): 23-27.

Xia Lixin. (2012). Putting Corpus Data into the Collocation Information in Chinese-English Dictionaries for Chinese Users. Foreign Language and Literature, (3): 47-50.

Yao Xiaoping. (2010). A Chinese-English Dictionary (Third Edition). Beijing: Foreign Language Teaching and Research Press.

THE PROBLEMS IN SELECTING *KBBI*'S ENTRY CANDIDATE FROM REGIONAL LEXICON

Dewi Khairiah & Dira Hildayani
National Agency for Language Development and Cultivation,
Ministry of Education and Culture of the Republic of Indonesia
dewikhairiah79@gmail.com
dirahildayani@gmail.com

Abstract

The variety of ethnic groups, culture, and languages in Indonesia has a great contribution to the enrichment of Indonesian vocabulary. *Kamus Besar Bahasa Indonesia* (KBBI) or The Great Dictionary of Indonesian Language as a descriptive dictionary has recorded the lexicons which come from both of the general languages and the regional languages in Indonesia as well as the foreign languages emerging from global contact. The online-based KBBI makes the task of KBBI compilers in collecting the data become easier since the KBBI users can contribute new entry directly via the online application. However, there are some problems related with the proposal of the lemma coming from the regional languages. In this paper, we describe the problems on lemma selection from the regional lexicons and the ways they solve the problems. For the purpose of this paper, we select 145 entry proposals sent from 29 regions in Indonesia as our primary data. The data is chosen based on the criteria defined in the general guidelines of Bahasa Indonesia's term formation (PUPI) and KBBI. The analysis shows that there are four problems faced by the KBBI editor in terms of regional lexicon proposals. The problems include (1) similar concept in Bahasa Indonesia (24%), (2) specific definition (8%), (3) non-standard grammatical form (17%), and (4) non-euphonic (51%). To sort the problems out, the editor takes two decisions: accepting or rejecting the entry proposals. Accepting requires modification or change in the form and spelling but should not cause change of meaning. Meanwhile, when an entry has similar concept in KBBI or cannot be modified into standard Bahasa Indonesia, it will be rejected.

Keywords: *entry selection, regional lexicon, lexicographic problems*

Introduction

Indonesian language, known as Bahasa Indonesia, was born on October 28, 1928, when the Indonesian youth from all over the country gathered in the Youth Congress to voice their pledge. This Youth Pledge or *Sumpah Pemuda*, declared three points: one motherland, one nation, and one language, that is Indonesia. At that time, Bahasa Indonesia was proclaimed to be a language of unity, a symbol of rebellion against the Dutch's colonialization. After Indonesia's independence, the status of Bahasa Indonesia was upheld as the official language of nation on August 18, 1945, as stated in the 1945 Constitution of the Republic of Indonesia, article 36.

Indonesia consists of a significant number of distinct ethnic groups that speak approximately 646 languages. With its diversity and great size of population, Indonesia needs a language that enables its people to communicate with each other (*lingua franca*). Malay language is chosen as Bahasa Indonesia because it has been used for hundred years in trade and it is easy to be learnt. As it develops, the vocabulary in Bahasa Indonesia is dominated by Malay, Javanese languages, and foreign languages that comes into Indonesia (especially Dutch and English). Unfortunately, from 700 languages in the archipelago, only 80 regional languages in Indonesia have contributed to enrich Bahasa Indonesia vocabulary, or approximately 4.45% of 108 thousands entries in KBBI. Although Bahasa Indonesia is used constitutionally (based on Act No. 24 of 2009) as a formal language in Indonesian people daily life, the position of vernacular languages or mother tongues is guaranteed to existence and development (Moeliono, *et.al.*, 2011).

Referring to the formulation of *Seminar Politik Bahasa* or Political Language Seminar (Alwi & Sugono, 2011), regional language is a *lingua franca* within an ethnic community alongside with Bahasa Indonesia, which supports local literature and culture. It is also an identity of an ethnic group and a valuable source for the enrichment of Bahasa Indonesia.

Background of the Study

Regional lexicons donate many new entries and enrich the expressions in Bahasa Indonesia. The launching of online version of *Kamus Besar Bahasa Indonesia* (*KBBI*), fifth edition, has enabled the penetration of regional lexicons into Bahasa Indonesia by giving access to crowd sourcing. This means that the KBBI's users can send their proposals of entry candidates to the editor team of KBBI through the online application. However, the proposals cause problems to the editors since the entries are not directly adoptable. There are some adjustments and considerations should be taken before they decide to include the entries into KBBI.

It is interesting to view how an editor chooses an entry and decides to forward it to the validator, the final decision maker who ratifies which entries should be recorded or not in KBBI. The decision is definitely based on the main principle, that is, the new entry is proper to be adopted into standard Bahasa Indonesia based on the criteria defined in KBBI and *Pedoman Umum Pembentukan Istilah* (*PUPI*) (Pusat Bahasa, 2005) or the General Guidelines of Term Forming. For example, a Sundanese verb *sisidueun* ("to hiccup") is not considered a suitable entry to borrow since it is difficult to be pronounced by most Indonesians outside the West Java area which speaks this language. Besides, KBBI has recorded another entry in Bahasa Indonesia whose conceptual definition is similar with *sisidueun*, it is *cegukan*.

There are several previous studies dealing with the donation of regional lexicons that enrich Bahasa Indonesia vocabulary. Two of them belong to Kulsum (2015) and Sudaryanto (2017). Kulsum (2015) describes the opportunity of some vocabularies in Sundanese language

to be adopted by Bahasa Indonesia by considering their meaning, category, and form. She found that some vocabularies are potential to be included into Bahasa Indonesia, especially those related to the plants names, kinship terms, and human body. Meanwhile, Sudaryanto (2017) made an inventory of Bahasa Indonesia vocabularies coming from lexicons of regional languages spread in Java. His research shows that there are five languages and dialects in Java which made contribution to Bahasa Indonesia vocabulary, they are Javanese (1,109 entries), Sundanese (223 entries), Madurese language (221 entries), Malay dialect of Jakarta (428 entries), and Using dialect (46 entries). From this result, it is obviously seen that Javanese language is the major donor for Bahasa Indonesia vocabulary.

Both studies talk about the regional lexicons adopted by Bahasa Indonesia. However, none of them deals with the acceptance process of entry candidates from the point of view of KBBI editor. In this paper, we discuss about the solutions taken in the selection of the problematic entry candidates based on our experience as KBBI editors.

Before being adopted as a new entry in standard Bahasa Indonesia, a lexicon from any languages (regional or foreign languages) should fulfill respectively all of the criteria defined in PUPI. First, the entry has unique concept. It means that the concept is not been recorded yet in Bahasa Indonesia. For example, a cultural term *tinggimini* from Muyu language in Papua is defined as 'the cutting of finger to show condolence for the death of one of family members'. There is no similar concept found in KBBI.

Second, the definition is neither so broad nor so detailed that represents a concept properly in Bahasa Indonesia. Take an example, the definition like "tree grows at the bank river" is not good because it gives a little information: how tall the tree is, how it looks like (the branches, the leaves, etc.). In contrast, too detailed or specific information is also not proper for a general dictionary's entry. For example, *ncue* from Kalisusu language (Southeast Sulawesi) means 'the number of house's steps' is a specific cultural terms suitable for regional language dictionary because its concept is not too significant for common Indonesian people to know. It is culturally related to a specific region.

Third, the entry should form or be formed grammatically following the standard language rule of Bahasa Indonesia in formal situation. For example, *adak* from Batak (North Sumatra) is the base to form verb *meng.adak* ('to consider her/himself cannot be fought back by others'). However, the application is not strict because in some cases, there are words that do not follow this rule. As an example, affix cannot attach to the verb *blusukan* since most users borrow it directly as it is from the source language (from Javanese language, means 'to come into a place to find out something'). As a result, it is categorized as informal in Bahasa Indonesia[1].

Fourth, it is euphonic or can be pronounced easily by common Indonesian people (as in the case of *kekeh* from *kekeuh* (Sundanese for 'persistent'). Besides, euphonic means that the entry consists of short words or syllables. Consider this entry from Kulawi language (Central Sulawesi): *tolumpole kaopompulukawu pulungkawu*, is it easy to memorize or spell it?

Fifth, it has good connotation and makes no ambiguity in meaning. For example, *lokalisasi* and *pelokalan* both are adoptable (from English *localization*), but the last term has better sense than the prior since *lokalisasi* is socially identified with *brothel*.

Sixth, it is widely and frequently used by a community for a long time. For example, *bobotoh* from Sundanese language (West Java) is a familiar term for most Indonesian people to call 'football supporters, especially from West Java area'. This term is used also in mass media.

This study is based on the criteria above to choose the regional lexicons which are not qualified and categorize them as problematic entry candidates. The criteria are also significant

[1] KBBI labels informal situation with *cak*, shortening for *cakapan*

as the basis to see the process of making decision in dealing with those problematic entry candidates.

Objective of the Study

Based on the background of the study, the main issues in this paper are the problems faced by the KBBI editor and how they are resolved.

Methodology

For the purpose of this paper, we selected 145 online entry proposals from 29 regions in Indonesia for our primary data. These entries were chosen based on the consideration that they were not followed the points required by KBBI and PUPI as good entries in Bahasa Indonesia. The data were categorized based on the problems we found, they were (1) having similar concept with KBBI, (2) having too specific definition, (3) having non-standard grammatical form, and (4) being non-euphonic. The analysis includes the description of the problems and the solutions taken by KBBI editor to deal with those problems.

Discussion

Based on the six criteria afore discussed, we selected 145 entry proposals which were not suitable and we compiled them into categorizations for analysis. Although PUPI defines sixth criteria, the data we collected only represent four criteria. Therefore, we categorized the data as follows: (1) having similar concept with KBBI, (2) having too specific definition, (3) having non-standard grammatical form, and (4) being non-euphonic. In this paper, we only take five data samples for each category to be analyzed.

Table of Problematic Entry Candidates

Having similar concept with KBBI	Having too specific definition	Having non-standard grammatical form	Being non-euphonic
baker (Alas language, Aceh)	*batuna gundu-gundu* (Wolio language, Southeast Sulawesi)	*beghukal* (Serawai language, Bengkulu)	*areuy geureung* (Sundanese, Banten)
kirik (Balinese, Bali)	*wakte* (Jambi Malay, Jambi)	*ga'ang payang* (Rejang language, Bengkulu)	*blunking* (Kerinci, Bengkulu)
carancang tihang (Sundanese, Banten)	*wan alang* (Alas language, Aceh)	*akeul* (Sundanese, West Java)	*mlalamdan* (Orya language, Papua)
kedhe (Javanese, Yogyakarta)	*anjong jahe* (Alas language, Aceh)	*bolitn* (Dayak language, Central Kalimantan)	*rompole katolompulungkawu* (Kulawi language, Central Sulawesi)
nandur (Jambi Malay, Jambi)	*ncue* (Kalisusu language, Southeast Sulawesi)	*marantikaq* (Benuak language, East Kalimantan)	*ngkawota* (Kalisusu language, Central Sulawesi)

The table above shows us four problems faced by KBBI editor with the entry candidates from regional lexicons. Here is the discussion about the problems and how KBBI editor sorts them out.

1. Having similar concept with KBBI

Five examples from this group are *baker*, *kirik*, *carancang tihang*, *kedhe*, and *nandur*. The concepts have been recorded in KBBI so they cannot be accommodated anymore in Bahasa Indonesia. *Baker* is similar with *migrain* ('migraine'), *kirik* with *kirik* (borrowed from Javanese language), *carancang tihang* with *syuruk* ('dawn'), *kedhe* with *kidal* ('to be left-handed'), and *nandur* with *tandur* ('to plant rice in the rice field'). Although the same concept cannot enter KBBI, some such entries are still possibly adopted on condition that they are widely used in mass media, books, or other publications. Take an example, kinship terms found in KBBI, such as *abang* (Malay, Betawi, and Kalimantan), *uda* (Minangese, West Sumatra), *mas* (Javanese), *akang* (Sundanese), and *daeng* (West Sulawesi) are similarly referred to 'a greeting name referred to older brother or respected man to show close relationship between the speakers'. The difference is that they are used based on the cultural background of the boy or the man to whom the name is referred. The entry candidates in which their concepts have been recorded in KBBI are rejected, unless their usage is broad among Indonesian society.

2. Having too specific definition

From the data we found, we chose *batuna gundu-gundu*, *wakte*, *wan alang*, *anjong jahe*, and *ncue* to be analyzed. As explained in the previous part of this paper, entry with too specific definition is closely related to cultural concept of a region. This kind of concept should be included into regional language dictionary rather than general Bahasa Indonesia dictionary. Specific cultural concept can be found in *batuna gundu-gundu*, a term for 'a stone used for initiation of leaders in Gundu-gundu community in Buton Island'. This community is not familiar among Indonesian people, thus the term is not suitable as Bahasa Indonesia vocabulary. Others represent kinship terms hierarchically, such as *wakte* ('greeting name referred to white-skinned siblings of our parents') and *wan alang* ('greeting name referred to the sixth brother of our father') or describe parts of a building in detail, like *anjong jahe* ('the southern sides of Alas community's traditional house') and *ncue* ('the number of steps of stairs in a house'). Those concepts are not too significant unless for the people living in the related area. Therefore, they are not required to be Bahasa Indonesia lexicons. The entry candidates with too specific definitions are rejected by KBBI editor.

3. Having non-standard grammatical form

From the table above, we can see the entry proposals which are not grammatically standard, they are *beghukal*, *ga'ang payang*, *akeul*, *bolitn*, and *marantikaq*. In Bahasa Indonesia, consonant cluster /gh/, as in *beghukal*, is not common, it will be modified into /g/. Thus, *beghukal* should be *begukal* if it is adopted into Bahasa Indonesia. Another word containing consonant cluster is *bolitn*. The KBBI editor should modify the last syllable [tn] by adding [ê], thus it will be *boliten* ([bOlitên]). Similar case happens to diphthong [eu] in *akeul* which should be modified into *akel* ([akêl]) before adopted by Bahasa Indonesia. Bahasa Indonesia also does not adopt glottal, like in *ga'ang payang*. To sort this out, the KBBI editor,

again, should modify the entry by deleting the glottal. Thus, the entry is changed into *gaang payang*. The last entry in this group is *marantikaq*, in which it contains uncommon last consonant /q/. In standard Bahasa Indonesia grammar, consonant /q/ should become /k/, so the entry should be *marantikak*. In summary, there is an opportunity for entries with non-standard grammatical form to be contained in Bahasa Indonesia vocabulary. However, it requires phonemic modification without changing meaning. Those which are not adaptable to the modification will be rejected.

4. Non-euphonic

Some of the entry proposals indicate non-euphonic or difficulty in the spelling. Five examples in this category are *areuy geureung*, *blunking*, *mlalamdan*, *rompole katolompulungkawu*, and *ngkawota*. *Areuy geureung* from Sundanese is difficult to be pronounced by non-Sundanese people because they are not used to the diphthong [eu]. This is different from entries that are included in non-standard grammar, non-euphonic group entries cannot be adopted into Indonesian if it is difficult to be pronounced. If editor should accept it into Bahasa Indonesia, it should be changed by deleting diphthong. Thus, it becomes *arey gereng.* Other entries have uncommon consonant clusters in Bahasa Indonesia, they are *blunking*, *mlalamdan*, and *ngkawota*. *Blunking* contains two consonant clusters /bl/ and /nk/ which are uncommon in Bahasa Indonesia since they are difficult to spell. Although /bl/ is still considered acceptable (some modern entries borrowed from foreign language have this cluster), cluster /nk/ should be inserted by /g/, so the entry becomes *blungking*. The same case happens to *mlalamdan*. Its consonant cluster /ml/ should be modified by vowel insertion, thus it becomes *melalamdan*. Meanwhile, since Bahasa Indonesia does not know consonant cluster /ng/ attached at the first syllable of a word, *ngkawota* should be initiated by vowel /e/. Therefore, it is changed into *engkawota*. However, it should be noted that the form change must not cause the meaning change. If it does, the entry proposal should be rejected.

Different case can be seen in the entry *rompole katolompulungkawu*. It consists of two words in which one of them is a long-syllable-word. It is not a good entry in Bahasa Indonesia because Indonesian people mostly have difficulty in memorizing or speaking long words. Therefore, this entry is possibly rejected, but it still depends on the validator's decision to accept it or not. Some considerations might be taken at the last, especially when the validator sees the significance of a term for Bahasa Indonesia in spite its criteria are not acquired.

Conclusion

Based on our analysis, we can conclude that there are four problems faced by the KBBI editor in terms of regional lexicon proposals. The problems include (1) similar concept in Bahasa Indonesia, (2) specific detail or definition, (3) non-standard grammatical form, and (4) non-euphonic. The decisions taken by KBBI editor to sort the problems out can be categorized into two: accepting and rejecting the entry proposals. Accepting requires modification or change in form and spelling but should not cause change of meaning. Meanwhile, the KBBI editor rejects the proposal when the entry has similar concept in KBBI or cannot be modified according to the standard Bahasa Indonesia. From 145 data proposals, we found that 24% entries have similar concepts with those in Bahasa Indonesia, 8% entries have too specific definitions, 17% entries have non-standard grammatical forms, and 51% are non-euphonic. Based on the percentages above, it indicates that most of regional lexicons are non-euphonic.

References

Pusat Bahasa. (2005). *Pedoman Umum Pembentukan Istilah* (3rd ed.). Jakarta: Pusat Bahasa, Departemen Pendidikan Nasional.

Kamus Besar Bahasa Indonesia. https://kbbi.kemdikbud.go.id/

Alwi, Hasan & Sugono, Dendy. (2011). *Politik bahasa: Rumusan Seminar Politik Bahasa*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.

Moeliono, Anton M., *et.al*. (2011). *Butir-butir perencanaan bahasa: Kumpulan makalah Dr. Hasan Alwi*. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.

Kulsum, Umi. (2015). Potensi bahasa Sunda dalam memperkaya bahasa Indonesia. *Bahasa dan sastra*, 15 (2).

Sudaryanto. (2017). Inventarisasi kosakata daerah bahasa Indonesia sebagai sarana konservasi bahasa: Kajian leksikologi. *The 6th University Research Colloquium*. Magelang: Universitas Muhammadiyah Magelang.

Building a Corpus-based Frequency Dictionary of Vietnamese

Dien Dinh[1], Nhung Nguyen Tuyet[2], Thuy Ho Hai[3]
[1,3]Computational Linguistics Center, Vietnam National University-HCMC
[2]Security University, HCMC-VN
[1]ddien@fit.hcmus.edu.vn, [2]velvetsnow.nguyen@gmail.com, [3]hohaithuy@hotmail.com

Abstract

In the age of Digital World, nowadays, corpus linguistics field has become more and more popular, bringing useful applications to the research of popular languages in the world. Based on corpora, corpus lexicographers can extract effectively linguistic features for each word, e.g. the word usage frequency, the popular POS (Part-Of-Speech), the grammatical pattern, the most often-used sense, etc. With reference to Vietnamese language, however, such applications are still limited due to the lack of available publicized corpora of Vietnamese. In this article, we present the exploitation of Vietnamese corpora, namely VfDic - Vietnamese frequency Dictionary, in building a full-fledged word-usage frequency dictionary of Vietnamese. From this dictionary, we extract the lists of the most popular vocabularies, e.g. the top-1000, top-2000, or top-3000 wordlist, accompanied with the most frequently used collocates (i.e nearby words), which provide valuable insight into the meaning and usage for learners of various learning levels. From these lists, we find out that: in Vietnamese, only 10% of the most popular words occupies 90% of the word tokens appearing in Vietnamese texts. This finding will be useful for teaching Vietnamese to foreigners, compiling definitions for entries in dictionaries, etc. Besides, this frequency dictionary can also be applied in the research of text readability, text stylometry and other NLP tasks of Vietnamese. In this work, we make use of an existing Vietnamese dictionary and two monolingual corpora, namely VCor (Vietnamese Corpus) and VTB (Vietnamese Tree Bank), provided by the Computational Linguistics Center (CLC, University of Sciences, Vietnam National University of HCMC. VCor consists of approx. 17M (million) sentences, 330M words, 440M morpho-syllables which have been word-segmented and POS-tagged automatically by CLC-tools. VTB consists of approx. 300K (thousand) sentences, 7M words which have been manually annotated with the Word Segmentation, POS, NER (Named Entity Recognition), etc.

Keywords: Frequency dictionary, corpus lexicography, corpus linguistics, Vietnamese language teaching.

1. Introduction

The word-usage frequency dictionary is one of the most engaging and efficient language resources for many practical applications, e.g. building the basic vocabulary sets, selecting approriate vocabulary sets for learners of different levels (Dinh, Kim, Nguyen, 2017), evaluating text readability (Luong, Nguyen, Dinh, 2017) in the textbooks, recruiting news reporters, determining text stylometry in the authorship attribution (Nguyen, Do, Dinh, 2018), etc. To build such a full-fledged word-usage dictionary, we need large corpora which cover various genres, domains, genders, regions, etc. These corpora must be respresentative and balanced, avoiding the gap or bias among domains or genres. Besides, those corpora need to be annotated with essential linguistic information, e.g. word boundaries, POS (parts-of-speech),… before being calculated with a wide range of statistic parameters, such as frequency of morpho-syllables, words, POS, etc. However, preparing corpora that meet all requirements for building a full-fledged frequency dictionary is a high-cost and time-consuming task. So far, there are no Vietnamese corpora publicized as such. Therefore, all the current frequency dictionaries of Vietnamese are based on the rank frequencies of morpho-syllables or words only (Dinh, Pham, Ngo, 2003; Pham, Patrick, Baayen, 2012). Meanwhile the frequency of word-usage based on its POS has been ignored completely. For example, the Vietnamese word "tốt" is used very frequently, but the frequencies of its collocates are significantly different. For instance, the frequency of "tốt" in the usage of "good" (adj) is very high, whereas the usage of "soldier" (noun) is very low. When selecting the most frequently used words for Vietnamese language teaching or text readablity evaluating, we must take into consideration their usages in terms of POS. It means that "tốt" (adj) should be selected as a frequently used word whereas "tốt" (noun) shouldn't be. This is a new and strong point of our frequency dictionary in comparison with other previous published frequency dictionaries of Vietnamese. To build such dictionary, in this project, we are licensed to access valuable corpora created by the CLC (2017). These corpora have been licensed to some organizations (e.g. I2R, Samsung, Systran, etc.).

In this paper, we present how to build a full-fledged word-usage frequency dictionary of Vietnamese from the aforementioned Vietnamese corpora. Beside the general introduction, the remainder of this paper consists of following sections:

- Description of Vietnamese corpora: collection, statistics, …
- Building a word-usage frequency dictionary: macrostructure, microstructure, …
- Results from the frequency dictionary: extracting the appropriate vocabulary sets for Vietnamese language teaching, measuring text readability and text stylometry, compiling lexicography.
- Discussion and conclusion

2. Description of Vietnamese corpora

In this work, we make use two corpora, namely VCor (Vietnamese Corpus) and VTB (Vietnamese Tree Bank):

2.1. VCor: this is an unannotated monolingual corpus collected from various sources: online news, books, … over a period of ten years, 2000 to 2010. This corpus consists of approx. 805K documents, 17M sentences, 346 M words and 440M morpho-syllables and covers 18 topics/domains. This corpus is automatically segmented in words by grouping all morpho-syllables within one word together, for example:

<S id='00001'> Chính_sách của Nhà_nước là đầu_tư xây_dựng nhà chung_cư bán cho người có thu_nhập thấp , nhưng rốt_cục lại không được quản_lý tốt </S>

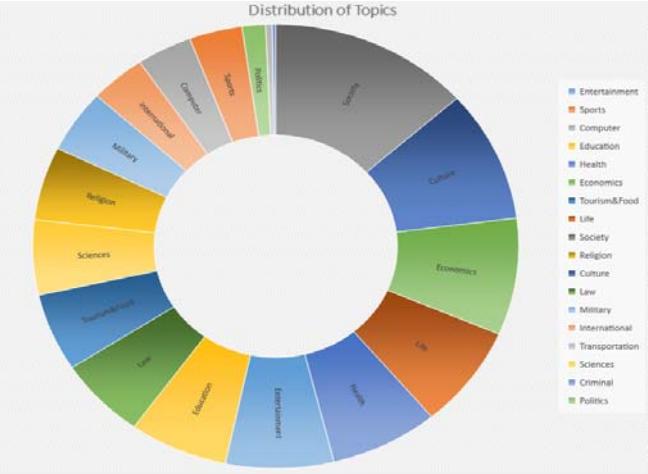


Figure 1. Distribution of Vietnamese corpus VCor.

No.	Topic	Ratio	Files	Sentences	Words	Mor-Syl	Lengt
1	Entertainment	7.19%	67,535	1,374,386	26,350,787	31,868,527	19.17
2	Sports	3.53%	32,945	668,776	12,609,716	15,660,217	18.85
3	Computer	3.70%	27,037	616,068	12,638,479	16,392,697	20.51
4	Education	6.57%	47,740	1,060,987	22,535,214	29,142,722	21.24
5	Health	7.23%	56,154	1,211,813	25,796,610	32,040,892	21.29
6	Economics	8.51%	55,360	1,284,164	27,840,867	37,715,850	21.68
7	Tourism&Food	5.65%	62,030	964,265	19,430,539	25,046,919	20.15
8	Life	7.45%	75,093	1,406,104	26,503,411	33,032,093	18.85
9	Society	13.39%	97,144	2,174,765	45,975,042	59,375,454	21.14
10	Religion	5.33%	39,320	942,721	18,984,779	23,618,434	20.14
11	Culture	9.56%	86,842	1,770,401	33,964,734	42,378,422	19.18
12	Law	5.90%	43,219	977,697	19,309,864	26,170,834	19.75
13	Military	4.58%	30,660	746,093	15,859,404	20,312,096	21.26
14	International	3.70%	27,073	595,851	12,506,045	16,418,458	20.99
15	Transportation	0.44%	2,811	66,352	1,563,769	1,958,420	23.57
16	Sciences	5.47%	44,035	954,725	18,496,247	24,234,407	19.37
17	Criminal	0.23%	2,328	41,419	736,881	1,019,988	17.79
18	Politics	1.56%	7,859	239,407	5,352,145	6,915,346	22.36
	Total	**100%**	**805,185**	**17,095,99**	**346,454,533**	**443,301,77**	**20.27**

Table 1. Distribution of Vietnamese corpus VCor.

The statistics from the total words and the average length of sentences ("word" is the orthography word). In Vietnamese, the majority (70%) is bisyllabic words, and the average length of a word is about 2.12 syllables (aka. morpho-syllable = orthography word).

2.2. VTB: this is an annotated monolingual corpus extracted from VCor. Its size is approx. 300K sentences, 7 million words which were manually annotated with linguistic information, e.g. Word Segmentation, POS (Part-Of-Speech), NER (Named Entity Recognition), etc., asshown in Figure 2 below.

```
<ANNOTATOR id="VTB0017">
 <DOC docid="V010973" Language="Vietnamese" Domain="News">
  <PARA id="1">
   <SEG id="1">Nguyên_nhân/Nn/O là/Vc/O bão/Nn/O số/Nn/O 10/An/O đang/R/O
chịu/Vv/O ảnh_hưởng/Nn/O bởi/Cp/O hệ_thống/Nn/O trục/Nn/O rãnh/Nn/O cao/Aa/O và/Cp/O
sự/Nc/O lôi_kéo/Vv/O từ/Cm/O siêu__bão/Nn/TRM_B Melor/Nr/TRM_I ở/Cm/O ngoài/Cm/O
khơi/Nn/O Philippines/Nr/LOC_B ./PU/O</SEG>
   <SEG id="2">Theo/Vv/O ông/Nn/TTL_B Bùi_Minh_Tăng/Nr/PER_B -/PU/O
giám_đốc/Nn/DES_B Trung_tâm/Nn/ORG_B Dự_báo/Vv/ORG_I khí_tượng/Nn/ORG_I
thuỷ_văn/Nn/ORG_I trung_ương/Aa/ORG_I ,/PU/O bão/Nn/O số/Nn/O 10/An/O có/Ve/O
hướng/Nn/O di_chuyển/Vv/O và/Cp/O diễn_biến/Vv/O rất/R/O phức_tạp/Aa/O ,/PU/O
có_thể/Aa/O thay_đổi/Vv/O so/Vv/O với/Cp/O nhận_định/Nn/O ban_đầu/Nn/O ./PU/O</SEG>
  </PARA>
 </DOC>
</ANNOTATOR>
```

Figure 2. A sample of VTB.

No.	Description	VCor	VTB
1	Number of sentences	17,095,994	302,491
2	Number of morpho-syllables	443,301,776	9,154,582
3	Number of words	346,454,533	7,096,580
4	Avg length of sentence (in words)	20.27	23.46
5	Avg length of word (in morpho-syl)	1.28	1.29
6	Avg length of morpho-syl (in letters)	3.27	3.27
7	Number of unique morpho-syls	6,835	6,714
8	Number of unique words	34,588	32,645

Table 2. Statistic of Vietnamese corpora Vcor and VTB.

3. Building the word-usage frequency dictionary

In building a dictionary in general or a word-usage frequency dictionary in particular, it is critical for us to solve the issues of macrostructure and microstructure in the dictionary. Macrostructure is the list of headwords which had been chosen under certain criteria. Since Vietnamese is a highly isolating language, the task of choosing Vietnamese headwords faces many challenges. As a matter of fact, there is an endless controversy over the definition or specification of Vietnamese word boundaries. For example: *đường thẳng* (line), *nhà tranh* (cottage), *nhà gạch* (brick house), *dưa hấu* (watermelon), *xe đạp* (bicycle),… are some among a great number of cases on which linguists haven't got a unanimous decision in terms of word boundaries. In this work, we follow the word-criteria in the general monolingual Vietnamese dictionary (Hoang, 1980).

In the other hand, microstructure is the internal structure of each entry, containing both linguistic and extra-linguistic information. In this dictionary, the microstructure comprises many fields of linguistic information (morphology, POS, grammar, semantics) and extra-linguistic information (frequency of word-usage according to its POS).

3.1. The macrostructure of dictionary

As mentioned above, we follow the criteria of word selection in Vietnamese dictionary by Hoang (1980). In addition, we also bring new criteria as follows:

3.1.1. The removal of classifiers

One of most distinguished features in Vietnamese is *classifier* which is absent in European languages (in some cases, a classifier is equivalent to the determiner/article "the" in English or "le, la" in French). Classifiers (or vice-nouns) are often used to specify the class of nouns. Each noun can only goes with its respective classifier(s), e.g. "sách" (book) often goes with classifiers "quyển" or "cuốn" as in "Đây là một **quyển** sách" (This is **a** book), not "Đây là một sách".

However, in other cases it stands alone, without any classifiers, as in "Tôi thích đọc sách" (I like reading books). So, in the macrostructure of Vietnamese MRD (machine readable dictionart), it is not advisable to include all these possible combinations "*quyển sách/cuốn sách*". This is the reason why classifiers will not be integrated in entries of our Vietnamese MRD. It means that in the macrostructure of our Vietnamese MRD, entries like "thư", "sách", "bò" are comprised instead of *bức thư/lá thư/cánh thư* (letter), *quyển sách/cuốn sách* (book), *con bò* (cow/ox), …

3.1.2. The use of word denoting categories

Unlike classifiers in Vietnamese, words denoting categories or subcategories will be integrated in the entry of dictionary, e.g. "máy" (machine) → *máy tính* (computer), *máy in* (printer), *máy quét* (scanner), *máy vẽ* (plotter), *máy phát* (generator), *máy đọc mã vạch* (bar code reader);"bộ"(device) → *bộ đếm* (counter), *bộ xử lý* (processor),etc. Regarding the words denoting categories, they may have high generality and popularity but are sometimes absent in use, we will note this feature in their microstructure, e.g. "bệnh" (disease) in *bệnh lao* (tuberculosis), *bệnh ho gà* (whooping cough), *bệnh uốn ván* (tetanus), etc.

3.1.3. The use of affixes

Similar to many inflecting languages (e.g. English, Russian, French, etc.), some Vietnamese words are formed by adding affixes, e.g. *-hoá* (-ize), *-viên* (-er/-or/-ist/-ian), *-học* (-ology), *bất-* (in-/non-/ab-), *liên-* (inter-), *siêu-* (meta-/super-/hyper-) as in *điện toán hoá* (computerize), *lập trình viên* (programmer), *tâm lý học* (psychology), *bất thường* (abnormal), *liên văn hóa* (intercultural), *siêu sao* (superstar). These derivations are formed by contrasting English derivational affixes and Vietnamese morphemes (which have Sino-Vietnamese origin).

In statistics, the macrostructure contains 38,300 headwords. Each headword is either mono-syllablic or multi-syllabic. The total number of unique syllables is 6,835 morpho-syllables which are pure Vietnamese syllables only, excluding loan syllables, ethnic syllables, e.g. biu, daklak, etc.

3.2. The microstructure of dictionary

The microstructure of this frequency dictionary of Vietnamese inherit from all the existing fields of linguistic information in an MRD which we had built formerly (Dinh, Pham, Ngo, 2003). We add only one new field, namely the word-usage frequency calculated from the aforementioned Vietnamese corpora.

3.2.1. Linguistic information

The linguistic information employed in this word-usage frequency dictionary of Vietnamese are listed as follows:

- Word form, e.g. "sách", "thắng lợi", etc.
- Word variations: lemmas, collocates, reduplicatives, etc.
- Parts-of-speech of word, e.g. noun, verb, adjective, etc.
- Subcategory: e.g. subcategories of nouns: countable nouns, uncountable nouns; subcategories of verbs: transitive verbs, intransitive verbs, …
- The meaning of word in English, e.g. "book", "win", …

3.3.2. Frequency rank

The frequent occurences of a word is measured by following formula: **Error! Objects cannot be created from editing field codes.**

where m is the number of occurrences and N is the length of corpus used for measuring. For example, f=3 means this word has occurred at the frequency 1/1000.

3.3.3. The statistic of VfDic:

Table 3 and Figure 3 both statistically and visually illustrate the distribution of the POS in VfDic respectively as follows.

No.	POS	Ratio	Qtt.	Num	Num
1	Noun	43%	16,302	2.03	7.95
2	Verb	28%	10,851	1.83	7.10
3	Adjective	20%	7,761	1.91	7.43
4	others	9%	3,386		
	Total	100%	38,300	2.12	

Table 3. Distribution of POS in VfDic.

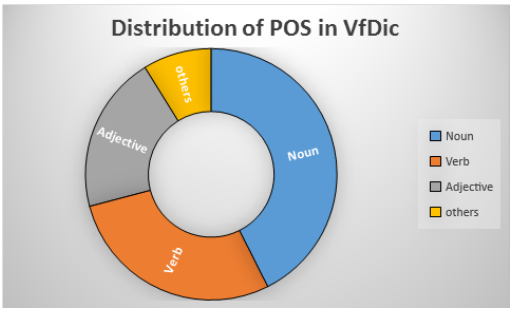


Figure 3. Distribution of POS in VfDic.

4. Results from the frequency dictionary of Vietnamese

We can search the word based on its POS. The statistics help us build the most frequently used vocabulary, e.g. top-1000 words, top-2000 words, or top-3000 words. These wordlist are suitable for learners of different levels. As shown in the Frequency of statistical results in Vietnamese, only 10% of the most frequently used words occupy 90% of the word tokens appearing in Vietnamese texts:

Rank	Word	Eng	POS	freq
1	của	of	Cm	1,820
2	và	and	Cp	1,822
3	các	+PL	Nq	1,956
4	có	have	Ve	1,959
5	là	tobe	Vc	1,968
6	trong	in	Cm	1,986
7	một	one	Nq	2,012
8	đã	+PST	R	2,031
9	những	+PL	Nq	2,043
10	không	not	R	2,050
..	…			..
14	người	man	Nn	2,160
25	nhiều	many	Aa	2,210
27	năm	year	Nt	2,314
30	ngày	day	Nt	2,401
31	làm	do	Vv	2,423
32	phải	must	Vv	2,436
34	ông	you	Nn	2,464
36	theo	follow	Vv	2,530
43	việc	thing	Nn	2,611
53	có thể	able	Vv	2,660

Rank	Wor	Eng	PO	freq
3,775	của	wealth	Nn	4,678
368	và	then	M	3,426
20,793	và	shovel	Vv	6,138
39,212	các	pay.extra	Vv	6,740
3,224	có	(particle)	M	4,573
103	có	exist	R	2,980
19,38	là	iron	Vv	6,041
5,290	là	being	Cs	4,920
143	là	as	Cp	3,085
1,749	là	(particle)	M	4,184
186	*tốt*	*good*	*Aa*	*3,181*
25,15	*tốt*	*soldier*	*Nn*	*6,439*

Table 4. Different word-usage frequencies

Legend: Cm: prepostion; Cp: conjunction; Nq: quantitier, Ve: V-exist; Vc: copula; R: adverb/adjunct; Nn: Noun, Vv: verb, Aa: adjective, M: Modifier.

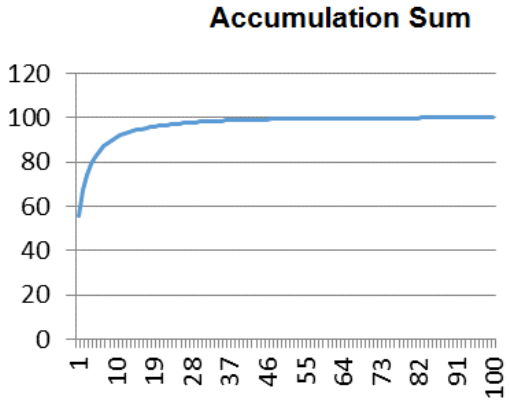


Figure 4. Accumulating Sum of VN words

Examples: the word "tốt" is used 3,624 times as an adjective ("good") but only 2 times as a noun ("soldier"). Similarly, the word "là" is used as a copula (tobe) much more often then other usages (as a verb, conj, prep, par).

5. Discussion and Conclusion

5.1. Discussion

Vietnamese is also similar to other languages as the most frequently used words belong to function words which are usually short in terms of length, e.g. "của, và, các" (in Vietnamese) and "the, of, and" (in English), "le/la, un/une, de" (in French), 的/de/(in Chinese), の/no/(in Japanese), etc. A morpho-syllable in Vietnamese is equivalent to a Chinese character (汉字 /hanzi/). Comparing the top-10 most common Vietnamese morpho-syllables with that of Chinese characters (Hua, 2018) and English (Wikipedia, 2018), we found that there are some similarities among them as shown in the table 5 below.

No.	Vietnamese	Chinese	English
1	của (of)	的 (of)	the
2	và (and)	一 (a)	be
3	các (+PLR)	了 (+PST)	to
4	có (have)	是 (be)	of
5	là (be)	我 (I)	and
6	trong (in, at)	不 (not)	a
7	một (a)	在 (at,in)	in
8	đã (+PST)	人 (man)	that
9	những (+PLR)	们 (+PLR)	have
10	không (not)	有 (have)	I

Table 5. Compare top-10 frequently used words in Vietnamese, Chinese and English

Regarding the cumulative sum in Figure 4, we found that the top 10% of Vietnamese word types occupies 90% of Vietnamese word tokens. There is a similar result in English. It is discovered that the top-3000 frequently used words also occupy 90% of all word tokens in English texts. This is also the reason why in the OALD8 (Oxford Advanced Learner's Dictionary 8th edition), all words that are used in the definitions of all headwords also appear within the top-3000 wordlist (Hornby, Dinh, 2014).

5.2 Conclusion

By exploiting Vietnamese corpora, we extract the most frequently used words, mostly based on their Part-of-Speech. This word-usage frequency dictionary can be applied to teaching Vietnamese language for foreign learners by building the wordlists that are suitable for learner of different levels. Besides, we can choose the appropriate words in compiling textbooks for learners who speaks Vietnamese as a foreign/second language and writing definitions in other dictionaries of Vietnamese. Last but not least, we can also exploit this word-usage frequency dictionary to measure the readability as well as stylometry of Vietnamese texts. In the future, if the word-usage frequency dictionary of Vietnamese is enhanced with semantic tags (from Vietnamese WordNet), its applications could be increased exponentially.

6. Acknowledgement

We would like to thank CLC (Computational Linguistics Center) for allowing us to access their corpora for building this word-usage frequency dictionary of Vietnamese.

References

CLC, (2017). Computational Linguistics Center, University of Science, Vietnam National University of HCMC http://www.clc.hcmus.edu., HCMC-VN.

Dinh, D., Pham, P. H., & Ngo, Q. H. (2003). *Some Lexical Issues in Electronic Vietnamese Dictionary*. PAPILLON-2003 Workshop on Multilingual Lexical Databases, Hokkaido University, Japan.

Dinh, D.,김위정, & Nguyen T. N. D. (2017). *Exploiting the Korean – Vietnamese Parallel Corpus in teaching Vietnamese for Koreans.* Interdisciplinary Study on Language Communication in Multicultural Society, the Int'l Conference of ISEAS/BUFS, 11-23.

Hoang, P. (1980). *Từ điển tiếng Việt*. Institute of Linguistics, Danang Publisher.

Hornby, A.S. & Dinh, D. (2014). *Oxford Advanced Learner's Dictionary 8th edition with Vietnamese translation*". Youth Publisher, HCMC.

Hua, S. L. (2018). Top 10 most Chinese characters. Retrieved from https://www.digmandarin.com/top-10-most-common-chinese-characters.html

Luong, A. V., Nguyen, T. N. D., & Dinh, D. (2017). *Examining the text-length factor in evaluating the readability of literary texts in Vietnamese textbooks*. 9th International Conference on Knowledge and Systems Engineering (KSE), Hue, 36-41. DOI: 10.1109/KSE.2017.8119431

Nguyen, T. N., Do, T. A. D., & Dinh, D. (2018). *Applying the text stylometry in detecting the gender of authors in Vietnamese texts.* The International Workshop on Vietnamese studies and Vietnamese linguistics, Hue.

Pham, H., Patrick, B., & Baayen, R. H. (2012). *Vietnamese word and syllabeme frequencies: a corpus and lexical decision study.* SEALS 22, France.

Wikipedia, (2018). Most common words in English: Oxford English Corpus. Retrieved from https://en.wikipedia.org/wiki/Most_common_words_in_English

Does the isiNdebele Terminology Developed Today Have Any Significant Impact?

Dr K.S Mahlangu, iZiko lesiHlathululi-mezwi sesiNdebele,
University of Pretoria, Pretoria, Republic of South Africa
sponono.mahlangu@up.ac.za

Abstract

After providing a brief background to the language policy in South Africa, the article explores the extent to which terminologies created by terminologists are understood and accepted by speech communities with a focus on isiNdebele. The isiNdebele language is one of the Nguni languages that is more under-sourced than the other official languages in the Republic of South Africa. The article will examine isiNdebele terminology and specifically some of the new terms found the Information and Communication Terms (2003), the Multilingual Mathematics Dictionary (2005) and the Multilingual Soccer Terminology (2009). Finally, the paper aims to provide some input on amendments regarding aspects of these terms that were coined in isiNdebele terminologies through compounding, for example.

Keywords: Compounding, Guidelines, IsiNdebele terminology, Multilingual terminology, Neologisms, Orthography, Standardisation, Terminology development

1. Introduction

The isiNdebele referred to in this paper is a South African Ndebele and not the Zimbabwean Ndebele. To provide for the recognition, implementation and furtherance of multilingualism in the Republic of South Africa and the development of previously marginalised languages the Pan South African Language Board was established. PanSALB established the National Language Bodies (NLB's) for all the South African languages, including the Southern Ndebele language. It also established the National Lexicography Units (NLU's) for all eleven official languages in South Africa. PanSALB regard the establishment and maintenance of the National Lexicography Units as the most important part of the primary comprehensive lexicographic process in South Africa. However, the main aim of establishing the NLB's was to address the issues of the development of orthography or spelling rules. IsiNdebele language is one of the Nguni languages that is less resourced than to other official languages in the Republic of South Africa.

2. Background

The dawn of the democratic South Africa in 1994, resulted in the nine indigenous African languages (viz. IsiZulu, IsiXhosa, SiSwati, IsiNdebele, Setswana, Sepedi, Sesotho, Tshivenda and Xitsonga) being accorded an equal status with Afrikaans and English. Hence, the new Constitution recognises eleven official languages. The isiNdebele language is the one recorded in the Constitution as one of the official languages in South Africa. This isiNdebele is in line with the constitutional designation, the so-called Southern Ndebele and not Northern Ndebele or Zimbabwean Ndebele (the Rhodesian Ndebele). It is one of the youngest languages to be offered official status. IsiNdebele as a standard form was encrypted in the year 1985. Language growth and development is one of the unavoidable types of behaviour of any language. Languages are never static; they are dynamic. IsiNdebele as a young language is going through a metabolic process of constant change. These changes also affect the formation of the coined words. Aitchison (2001:249) opines that a language gradually transforms itself and it cannot remain unaltered; thus isiNdebele is faced with the influx of newly coined words. These words come from the natural sciences, as well as mathematical, technological, HIV and AIDS terms derived from foreign languages, especially Greek, Latin, English and Afrikaans. Since 1994, the Department of Arts, Culture, Science and Technology (DACST) has employed African language terminologists to develop and document African language terminology in a variety of subject fields (isiNdebele included).

In the past, the National Terminology Services (NTS) used to work in collaboration with the old Language Boards and currently the terminologists of the National Language Service (NLS) work in consultation with the National Language Bodies. After a specific terminology list has been finalised, it is then taken to the language bodies for verification and authentication of terminology, so that they can also assist with the standardisation and stabilisation of terms as well as with popularising newly coined terms. Previously isiNdebele had only one terminology book that was published in 2001, i.e. The isiNdebele Terminology and Orthography No.1 Book of 2001. Van Huyssteen's (2003:238) advice during terminology development was that terminologists should take notice of popular phonological trends in the language and where needed spelling should be adjusted to suit the phonology of language change. So, this terminology book serves as the term bank for isiNdebele terms as it is the only terminology book that was issued by the Department of Education and Training. What is noted in the terms themselves is that a number of the isiNdebele terms have been adopted and *ndebelised* rather than coining new terms for terms equivalent to English and Afrikaans.

To-date the iKhwezi NLB, i.e the name of the isiNdebele Language Body had already approved seven lists before they were used, namely, the Mathematical term list (September 2005), OBE (Outcomes Based Education) terms (May 2002), HIV and Aids term list (July

2002), Information and Communication (ICT) terms (July 2003), Parliamentary / Political term list (2008), Multilingual Natural Sciences terms (2008), Multilingual Soccer term list (2010). There is a need for the isiNdebele terminologist to coin indigenous isiNdebele terms instead of relying on the adoptive form. The thrust that appears to be most prevalent is the one of using strategies such as semantic shift or transfer, borrowing, derivation, neologisms, compounding, deideophonisation and paraphrasing or phrase grouping.

Terminology is a phenomenon of specialised subject areas, which is also influenced by the subject fields and the areas of activity it serves. An isiNdebele terminology development in a variety of subject areas has been compiled by the Department of Arts and Culture, including, weather terms, basic health terms, HIV/AIDS terms, building terms, election terms, banking terms, commercial and financial terms, computer terms, mathematical terms, natural science terms, soccer terms, and water and sewerage terms. These have been completed but are not properly disseminated to the speakers of the language. Consultation is important when coining terms. Subject specialists, linguists; mother-tongue speakers and language committees such as language boards must be consulted when providing term equivalents or when coining terms. Should this process be bypassed, then the terminologies compiled will be ignored by the speakers, because terminologists have worked in isolation.

3. Objectives of the study

To illustrate how some of the new words have been adapted/adopted into the lexicon of isiNdebele, which like any living language is subject to constant change.

4. What is terminology development?

Terminology development is an area of focus that is imperative for accurate communication in technical fields and terms are created naturally and on an ad hoc basis (Masasanya 2005:8-10). Batibo (2009:14) maintains that terminology development is concerned with the creation, recording and institutionalising of lexical items. Osborn's (2010:41) view on terminology focuses on language change and planning that will be particularly relevant to localisation. He adds that in most cases terminology planning is informed by the new domains of language use, the level of suitability of terms in a given domain, policy and decision-making, plus the implementation strategies as well as the evaluation of capacity and extent of usage (Batibo 2009:14).

Van Huyssteen (2003:58) argues that the development of terminology in African languages (isiNdebele included) is unfortunately characterised by compilers having little knowledge of the theory of term development and also a lack of documented terms. Mnguni (2004:7) adds that African languages are faced with a serious challenge in as far as term creation in technology is concerned.

5. Terms coined by the isiNdebele Terminologists

The following are examples of some of the terms that were coined by the terminologists.

(a) Ingaphakathingqondomtjhini ***'software'***. It consists of the following concepts: *nga-* 'potential' *phakathi* 'inside' (adverb of place), *umtjhini* 'machine' (noun loaned from English), *ingqondo* 'mind' (noun)

(b) Ithungelelwanohlanganiso ***'internet'*** *thungelela* is 'to light fire' (verb), -an- (reciprocal extension) and *–hlanganisa* 'to combine' (verb)

(c)umbiko-mthethokambiso ***'white paper report'*** a 'white paper' consisting of a noun *umbiko* 'report', noun *umthetho* 'law' and a noun *ikambiso* 'system';

(c) Isibonisi-sidlalisimdumo ***'video cassette recorder'*** is also a compound consisting of the deverbative *isibonisi*, derived from the verb *-bonisa* and the deverbative *isidlalisi*, derived from the verb *-dlala* and a noun *umdumo* 'sound'

From the examples of terms created it shows that the terminologists ignored the advice of Sager (1990) and Taljard (2008) of adhering to the guidelines that should be used during term creation. To avoid some of the inaccuracies in the linguistic formulation of orthographical rules, Thipa (1989:180) and Mathumba (1993:210) are correct when they opine that Language Boards (currently known as National Language Bodies) should be changed to include more members who are knowledgeable and qualified in linguistics and language planning. If this situation were to be improved, this would enrich terminological development in isiNdebele.

6. Challenges facing the coined terms

If one looks at the above mentioned terms, there are some orthographical problems that are identifiable in as far as a hyphen is concerned. Some compounds are hyphenated and others are unhyphenated without any given reason(s) and this causes confusion and inconsistencies. On the one hand, it is apparent that coinage is still problematic in isiNdebele because, when coining terms, terminologists are trying to bring forth all the resemblances that are found in the Source language (SL) term.

7. Terms preferred by the isiNdebele speakers

Alberts (1999:28) points out that transliteration and borrowing develop the language, and terms can be coined according to transliteration principles. In most cases, the amaNdebele speakers prefer transliterated lexical items to the coined lexical items, because users opt for words which are closer to the source language (English / Afrikaans) and which have meanings equivalent to the original foreign items.

- Terms should be brief and they should not contain unnecessary information. In isiNdebele most of the terms that are short are the transliterated terms, e.g., *umrhatjho* 'radio' referring to the transmission and reception of radio waves especially those carrying audio messages. This term is short and meaningful because that is exactly what *umrhatjho* is in isiNdebele. In contrast, a term such as *isikhadlanammumatho* 'byte' for a unit of memory size of a computer, contravenes the guideline, because it is long and contains unnecessary information.
- Terms should be self-explanatory and transparent. In isiNdebele there are terms that are self-explanatory as well as those that are confusing. On the one hand, terms such as *ikhomphyutha* 'computer' and *umthumeli* 'sender' are self-explanatory and transparent, while on the other hand, terms such as *isikhadlanammumatho* 'byte' and *ingaphakathingqondomtjhini* 'software' are not self-explanatory and transparent because their meaning can be confusing, as already argued.

8. Transliterated lexical items

The transliterated lexical items are usually characterised by having the same meaning as their foreign counterparts.

Examples of transliterated lexical items extracted from the Mathematical term list (2005):

igrafu 'graph'
i-abhakhasi 'abacus'
idayamitha 'diameter'

Examples of transliterated lexical items extracted from the *Multilingual Information and Communication (ICT) terms* (2003):

i-imeyili 'e-mail'
iseva 'server'
idatha 'data'

From the above examples it becomes clear that, in isiNdebele, the transliterated lexical items are usually characterised by having the same meaning as their foreign or source counterparts and are understood by the speakers. The new terms such as *iselula* 'cell phone', *ikhomphyutha* 'computer' and i-*imeyili* 'e-mail' are new example terms that have been coined due to the growing technology across the globe.

9. Creation strategies

The article is in agreement with the assertion by Sibula (2009:87) and Alberts (2013:40) that terms should not be created haphazardly as there are specific ways of supplying term equivalents. Thus, terminologists should take cognisance of the various creation strategies such as the ones below when they coin terms in isiNdebele:

(i) Paraphrasing

This is another productive method of word formation that is used in isiNdebele and it occurs when new terms are created by a translation of the meaning of a foreign term into isiNdebele which is a target language. The following examples and their equivalents are extracted from the *Multilingual Mathematics Dictionary* (2005):

English 'currency' is translated as - *irherho lemali 'a system of money'*
English 'decade' is translated as -*itjhumi leminyaka 'ten years'*
English 'doubling' is translated as -*ukubuyelela kabili 'to repeat twice'*

(ii) Shortening

This is a process by which a word or words are omitted, usually as will be supplied instinctively or will be taken for granted as understood in the construction of a sentence. This is another creation strategy where a word is omitted from a compound expression of a source language but the remaining part still retains the total meaning that formerly belonged to the whole expression (Louwrens 1993:10). The isiNdebele word '*ikondasi'* is an adaptation of the English compound word 'condensed milk'. Because of the omission of the second part of the English word 'milk' the word '*ikondasi'* carries the meaning 'milk' as part of the concept process of condensation. The word '*iselula'* is also an adaptation of the English compound word 'cell phone'. Again because of the elision of the second part of the English word 'phone' the word '*iselula'* carries the meaning 'phone'.

(iii) Deidiophonisation

Apart from the above mentioned creation strategies, Mtintsilana and Morris (1988:111) contend that deideophonisation is another creation strategy. Moreover, van Huyssteen (2003:114) views this creation strategy as a unique method of word formation found in African languages. This strategy is termed because of the relationship between concepts to symbol. For example, *isithuthuthu* 'motorcycle' is one of the ideophonisation terms that has been coined in isiNdebele. Another example is *ithothotho* 'man-made traditional beer'. A prefix *'i-'* is added to a prototypical perception of the sound made during the process when the distilled beer forms some droplets and those drops as they enter into a container to become beer make a dripping sound *'tho…tho…tho'*; thus, this liquor is called *ithothotho*.

(iv)Borrowing/loan words

This is a process whereby words are loaned or borrowed as they are (wholes) and their meanings have remained as they are; they exhibit a varying degree of adaptation on the syntactic, morphological, topological and phonological levels (Louwrens, 1993:9). In isiNdebele, this type of borrowing takes place mainly from English, Afrikaans and Sotho languages as these are the languages of common contact. The English term 'virus' is equivalent to *ivayirasi* in isiNdebele and it conforms to the word forming principles of isiNdebele as it consists of a CVCVCVC structure.

(v) Extending or widening

This is a process when the meaning of the existing lexical item is extended to refer also to the meaning of the new term. The English word 'steamer' is regarded as 'steam engine train, diesel/electric train' and in isiNdebele *isitimela* meaning steamer is stretched to other types of trains whereas in its source language this word means 'steam engine train' alone.

Recommendations

The study recommends that when coining terms they should be short and to the point and they should be meaningful and also be understood by the speakers. If terminologists are unable to coin a term that is self-explanatory and transparent, they should consider coinage through transliteration because the transliterated terms are not only self-explanatory but they are also short and to the point. See the examples above (i)-(v).

Alberts (2013:45) advises that when coining term equivalents to Source Language (SL) terms, subject specialists, linguists, mother tongue speakers and language communities should be consulted, because consensus must be reached as subject specialists know the subject or domain and linguists could give authority to the term equivalents. If coinage is done following this advice then most of the coined terms will carry authority and may therefore be more readily accepted by the community of speakers. Should this process of consultation be bypassed, then the terminologies compiled will be ignored by the speakers, because terminologists have worked in isolation. The reason for ignoring these terms is that they were never maintained and popularised or maintained among the speakers. What is notable in isiNdebele is that these terms were not evaluated in terms of acceptability by the speakers; thus in most cases the speakers of isiNdebele end up preferring the transliterated terms to the newly coined terms.

References

Aitchison, J. 2001. *Language Change : Progress or Decay* ? Third Edition: Cambridge University Press.

Alberts, M. 1999 Theoretical principles of terminology and Terminography. Tutorial on Principles, Procedures and Practice of Terminology presented by the African Association for Lexicography (AFRILEX) at the University of Pretoria, Economic and Management Sciences Building, 29-30 November 1999. Notes 1-28.

Alberts, M. 2013. Developing Legal Terminology in African Languages as Aid to the Court Interpreter: A South African Perspective. *Lexikos,* 23: 29-58.

Batibo, H. 2009. Language Documentation as a Strategy for the Empowerment of the Minority Languages of Africa. *In Selected Proceedings of the 38th Annual Conference on African Linguistics*, ed. Masangu Matondo, Fiona Mc Laughlin, and Eric Postdam, 193-203. Somerville, MA: Cascadilla Proceedings Project.

Department of Arts and Culture. 2002. Multilingual Natural Science and Terminology Dictionary. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2003. Multilingual Information and Communication (ICT) terms. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2006. Multilingual Mathematical Dictionary. Pretoria: Department of Arts and Culture.

Louwrens, L.J. 1993. Semantic change in loan words. *South African Journal of African Languages*, 13(1):8-16.

Masasanya, B.D. 2005. Terminology usage in Setswana radio and Television. Unpublished M.A. dissertation, Johannesburg: University of Witwatersrand.

Mathumba, D.I. 1993. A comparative study of selected phonetic, phonological and lexical aspects of some major dialects of Tsonga in the Republic of South Africa, and their impact on the standard language. Unpublished D Litt et Phil thesis, Pretoria: University of South Africa.

Mtintsilana P.N &.Morris, R. 1988: Terminography in African languages in a South Africa. *South African Journal of African Languages* 8(4): 109-113.

Osborn, D. 2010. *African Languages in a digital age*: *Challenges and opportunities for indigenous language computing*. Pretoria: HSRC Press.

Sager, J.C. 1990. *A Practical Course in Terminology Processing*. Amsterdam/ Philadelphia: John Benjamins.

Sibula, P.M. n.d. Terminology Development: Challenges, skills, strategies and stakeholders' role. Unpublished Paper. Unit for isiXhosa, Stellenbosch: Stellenbosch University Language Centre.

Taljard, E. 2008. *Issues in Scientific terminology in African Languages*. In M. Lafon, &V. Webb (eds.) *IFAS* Working Paper Series, 88-92.

Thipa, H.M. 1989. 'The difference between rural and urban Xhosa varieties: sociolinguistic study.' D. Phil. Thesis. Pietermaritzburg: University of Natal.

Van Huyssteen, L. 2003. A practical approach to the standardization and elaboration of Zulu as a technical language. Unpublished D Litt.et Phil thesis. Pretoria: University of South Africa.

Collecting Etymological Information of Indonesian Malay Lexicon from Diachronic Corpora

Dewi Puspita
University of Indonesia
dewi_puspita18@yahoo.com

Abstract

Indonesian lexicon comprises numerous loanwords which some of them already exist since the 7th century. Loanwords in Indonesian come from Sanskrit, Chinese, Arabic, Persian, Portuguese, Dutch, and English. The large number of loanwords is the reason why many dictionaries of Indonesian etymology available today contain merely the origin of the words: where the words come from and the original form of the words. Some Indonesian etymology dictionaries even present one language origin only, like *Arabic loanwords in Indonesian* (Jones, 1978). Meanwhile, there are many things in a word etymology that can be studied and presented in a dictionary, such as the change in word form and change in meaning (Durkin, 2009). Moreover, Indonesian lexicon also has its roots: Malay, whose etymology also need to be revealed. Diachronic corpora can be a useful tool to investigate stages of word or language change. Fortunately, there are a number of collection of old Malay manuscripts from the 14th to the 20th century AD, that have been transcribed by philologists and compiled into a corpus by a project at the *Australian National University* called *Malay Concordance Project* (MCP) (Proudfoot, 1991). By using data from the MCP together with other corpora set diachronically, we can collect etymological information of Indonesian lexicon originated from Malay. We can also find out what kind of changes the words undergo from time to time. In this study, the words to be traced etymologically are *bersiram* and *peraduan*. Those words are native high-Malay words that are still used in Indonesia until today. By using diachronic corpora and applying concordance and collocation analysis method, this study attempt to collect as many etymological information of the words as possible. The results of this study can be used as data for a more comprehensive Indonesian etymology dictionary.

Keywords: diachronic corpora, etymology, *bersiram*, *peraduan*

I. Introduction

1.1 Background of the study

For most language users, etymological information is just information of where a word originated from. Especially when the language has so many loanwords, like Bahasa Indonesia. Kridalaksana (2001) states that the content of Indonesian etymology dictionaries which have been compiled and available today is merely an inventory of words origin which needs to be continued with research and interpretation from various aspects. This is in line with the opinion of Durkin (2009) and Liberman (2009) which state that the etymology study is related to the history of a word, the history of meaning, formal history, or the history of its spread from one language to another, or from one group to another. There are at least six etymological information that can be traced from a word: the year of usage; the initial form (morphology); the initial sound (phonology); the language of the donor (for loan word); the person who coined the word for the first time; the initial meaning and the change of meaning. Therefore, an etymological dictionary should not only contain information of the word's origin.

Other things from Indonesian etymology dictionaries that available until today is the absence of etymology of words originated from Malay. Malay is the root of Bahasa Indonesia. In the early centuries, the language spoken in some part of the Indonesian archipelago and the Malay Peninsula might be the same. Over time, there are many things and events, socially and politically that affect the regions and cause the language to change and to be different. Information of changes that occur in Malay words that now become the vocabulary of Indonesia, phonologically, morphologically, semantically, or syntactically are parts of etymology information.

To find out what changes a word has gone through, a tool is needed. The right tool that can provide a large collection of text from past centuries to be examined is diachronic corpora. According to Allan and Robinson (2012), the use of corpus is the state of the art in the study of historical semantics, which is part of etymology study. Malay is lucky to have *Malay Concordance Project* developed by Australian National University (Proudfoot, 1991). It consists of old classical Malay manuscript from 14th to 20th centuries that can be used to trace the usage of a Malay word throughout that time. Employing the Malay Concordance Project compared with a more recent corpus from the 21st century, this study will search for any etymology information of two Malay words that become part of Indonesian lexicon and still used until today. The words to be investigated are *bersiram* and *peraduan*. Those words are classical and used only to refer to royal family. Since now there are not so many royal families in Indonesia, there is a possibility that the usage or the meaning of the words might change.

1.2 Objectives

The objectives of the study are as follows.

1. To find any etymological information of Malay words *bersiram* and *peraduan* from diachronic corpora.
2. To investigate what kind of changes those Malay words undergo from time to time until they become Indonesian lexicon.

1.3 Methods

To prove that etymological information can be collected from diachronic corpora, this study will use two corpora that are set in chronological order. The first corpus is *Malay Concordance Project*, which comprises 5.7 million words (including 130,000 verses) from more than 150 sources of pre-modern Malay written text. The oldest script is from the year 1302 and the youngest is from 1950. However, the dates of some old scripts are somewhat

hypothetical. The second corpus is Indonesian corpus from *Leipzig Corpora*. This corpus is based on online material from 2012 to 2014 that consists of 74,329,815 sentences, 7,964,109 types, and 1,206,281,985 tokens. The two corpora will present the usage of Malay lexicon from the 14th to 21st century.

The search results of the words investigated from the two corpora will then be analyzed qualitatively. The changes that each word undergo will be examined from the concordance lines and the word's collocations. Collocation analysis usually involved statistical measurement. Yet McEnery and Hardie (2012) proposed a non-statistical method called collocation-via-concordance technique. In this technique, researchers must use their intuitive to scan the concordance lines that yields up notable examples and patterns and examine each line individually. This technique will be employed in this study.

II. Results

2.1. *bersiram*

The word *bersiram* is a high, classical Malay word. The word has been recorded in the dictionaries of Malay (Kamus Dewan, 2015) and Indonesian (Kamus Besar Bahasa Indonesia, 2016) with the meaning of 'to take a bath'. The word can only be used for the royal family. In Malay Concordance Project (MCP), the word appeared 157 times in 24 old scripts dated from the year the 1370s to 1930s. All those 157 tokens of *bersiram* in the contexts show the same meaning with those recorded in dictionaries. Below are some examples of the word in contexts:

1370s

(1) *sudah Élah kembali itu, maka baginda pun pergilah* ***bersiram*** *ke kolam itu. Setelah sudah baginda bersiram itu,*

1770s

(2) *Setelah selesailah daripada bercukur dan* ***bersiram*** *putera Baginda itu, maka datanglah bidan menjunjung duli …*

1810s

(3) *… anéka jenis daripada bungaan. Setelah sudah mandi* ***bersiram*** *maka naiklah segala puteri-puteri itu mengentas bunga2an ada yang …*

1890s

(4) *… sama elok parasnya. | Setelah genap tujuh hari,* ***Bersiramlah*** *baginda laki isteri, Dikerjakan oleh perdana menteri,*

1910s

(5) *Pada suatu hari Sultan Mahmud hendak berangkat* ***bersiram****, duduk di atas julangan, ditikam oleh Megat Sri Rama dengan …*

1930s

(6) *... bestari, manakala siang keluar matahari, selesai* ***bersiram*** *mahkota negeri. | Berangkat keluar ia bertakhta, tersenyum …*

The above concordance lines show that the *word* bersiram collocated with the word *baginda* (king), *puteri-puteri* (princesses), *perdana menteri* (prime minister), Sultan Mahmud (King Mahmud), and *mahkota negeri* (crowned head). The other concordance lines which are not presented here also show the same collocates. Those collocates indicate that the word *bersiram* is only used for the royal family. The line from the 1890s (sentence number (4)) even shows that the bath was not just a usual bath, it was a kind of ceremony.

(4) *... Setelah genap tujuh hari,* ***Bersiramlah*** *baginda laki istri, Dikerjakan oleh perdana menteri, ...*

... After seven days, The king and his queen ***took a bath***, Done by the prime minister, ...

After its independence in 1945, Indonesia has become a republic. The royal system is no longer used. For that reason, the frequency of use of the word *bersiram* might also be

decreased. However, in a more recent corpus like Indonesian corpus in *Leipzig Corpora*, we can still find the use of the word *bersiram* in many different contexts. The search of the word *bersiram* in Leipzig Corpora returned in 55 lines. There are many interesting things found from the lines:

a. From 55 occurrences, only 15 of them have the literal meaning of 'to take a bath' or 'to shower'. Ten lines, which come from Malaysian website, use the word *bersiram* as the equivalent of to take a bath or shower in the daily activity of common people, while the other five lines, which come from Indonesian website, still use the word only for a respected person.

b. Six lines contain the word *bersiram* in a figurative meaning. It collocates with *darah* (blood) and *cahaya* (light) as in the sentence (7) below:

(7) *Di kejauhan tampak gedung-gedung jangkung yang* ***bersiram*** *cahaya lampu.*
(mayasanti.blogspot. com, *crawled on 08/05/2012*)
In the distance, tall buildings are seen ***bathed*** in light.

c. The most interesting thing is, 34 lines of them appeared in the contexts of food and carry a figurative meaning. In those lines, *bersiram* mostly collocates with *saus* (sauce), *jamur* (mushroom), *keju* (cheese), *cokelat* (chocolate). One example of the word usage in the context of food is as follows:

(8) *Dari deretan menu terbaru, ada BBQ Beef Ribs & Alice Springs Chicken* ***bersiram*** *saus keju Monterey Jack-Cheddar*. (www.femina.co.id, crawled on 06/02/2014)
From the latest menu, there are a BBQ Beef Ribs & Alice Springs Chicken ***covered*** with Monterey Jack-Cheddar cheese sauce.

We can see from the two diachronic corpora that there are changes in the meaning of the word *bersiram*. The word that originally had only one meaning and used only for certain circle, after the twentieth century its meaning has widened to a figurative meaning, and move from specific to a more general meaning.

It is not only the semantic aspect of the word *bersiram* that change over time. Another linguistic aspect that also changes is the syntactic aspect. *Bersiram* is an intransitive verb by nature. In Indonesian grammar, prefix *ber-* forms intransitive verb. As can be seen in the sentence (1):

(1) *... maka baginda pun pergilah* ***bersiram*** *ke kolam itu.*
… then the King went to the pool ***to take a bath.***

The phrase *ke kolam itu* in above sentence is not an object, it is an adverb of place. An object is not needed after the word *bersiram* in that sentence.

However, in its figurative meaning, the verb *bersiram* has become transitive. Below is a concordance line of the verb *bersiram* in figurative meaning followed by its objects (in upright letters).

(10) *Tempat orang berniaga dikepalai seorang batin bijaksana yang mengharamkan negeri* ***bersiram*** **darah**.
(11) *Di kejauhan tampak gedung-gedung jangkung yang* ***bersiram*** **cahaya lampu**.
(12) *Sejumput mi* ***bersiram*** **saus** *dengan potongan udang gemuk di atasnya.*
(13) *Versi Michel's disebut Marble Mud Cake,* ***bersiram*** **ganache** cokelat putih dan cokelat pekat.
(14) *Dan, sebagai penutup pesanlah Roti Cane Gula atau Roti Cane Susu,* ***bersiram*** **susu** *kental manis.*

Objects in above sentences are mandatory because without objects the sentences would be incomplete and meaningless.

2.2. *peraduan*

The same as *bersiram*, *peraduan* is also a classical, high Malay word that is used only among the royal family. It has the meaning of 'bed' or 'bedroom'. The frequency of the

word's appearance in Malay Concordance Project is quite high. It appeared 357 times in 31 old manuscripts dated from the 1370s to 1950s. Here are some example from the concordance lines, all with the meaning of 'bed' and 'bedroom'.

(15) *.. ketiganya itu pun masing-masing mendapatkan biliknya* ***peraduan****, lalu beradulah sekaliannya itu.*

(16) *... beri rawan, sendu rupa kelakuan, buka ranjang* ***peraduan****. | Lalu makai Sinyor Gilang, baju lakan hitam gilang*

(17) *Maka Kuda Nestapa pun masuk ke dalam* ***peraduan*** *lalu menyingkap tirai kelambu itu. Maka dilihatnya Raden…*

(18) *Sambil memakai bau-bauan. Adinda disambut masuk* ***peraduan****. | Lalulah duduk menanggalkan jubah,*

(19) *... biliknya dan pada tiap-tiap bilik itu ditaruhnya geta* ***peraduan*** *lengkap dengan kasur, tilam dan tirai ...*

(20) *... bilik yang indah. | Istana besar apa gunanya,* ***Peraduan*** *lengkap dengan perhiasannya, Asingnya tidak ...*

Translation to number (15) and (17):

(15) All three were each got their own ***bedroom***, then they went to bed.

(17) So Kuda Nestapa went into ***the bed*** and unveiled the curtain. …

In a more recent corpus such as Leipzig Corpora, the frequency of occurrence of the word *peraduan* is also high. There are 650 occurrences from websites dated from 2012 to 2014. However, the meaning that the word carries in this corpus is rather different from those in Malay Concordance Project. From about 100 lines examined from the concordance lines, there are three types of usage of the word *peraduan*.

The first type has the same meaning and usage as those in previous corpus, which is bed or bedroom of the royal family. The word *peraduan* in the first type, as shown in sentences number (21), (22), and (23), are collocated with *raja* (king) and *kerajaan* (royal).

(21) *Sementara itu, sang raja telah tidur di* ***peraduan*** *kerajaan.*
Meanwhile, the King had slept in the royal ***bed***.

(22) *Jika nanti sudah berada dalam* ***peraduan*** *raja, cincin itu harus dilepas, dan ditaruh didekat Pusaka Keraton karena dirinya sudah berada di dalam cincin itu.*
When already in the king's ***bedroom***, the ring must be taken off and placed near the heritage of the palace because he is already in the ring.

(23) *Sebelum mencabut tombak, ia kembali keluar dari* ***peraduan*** *raja yang kesakitan itu.*
Before pulling the spear, he came back out of the afflicted king's ***bedroom***.

In the second type of usage, the word *peraduan* carries the same meaning but it is used by common people.

(24) *Membaca buku, majalah, atau sekadar mendengarkan musik, sebelum Anda beranjak ke* ***peraduan*** *untuk tidur.*
Read book, magazine, or simply listen to the music before you go to ***bed***.

(25) *Pagi itu hujan deras menguyur kota Surabaya dan sekitarnya, membuat badan malas untuk bangkit dari* ***peraduan****.*
That morning, heavy rain was pouring in Surabaya and its surrounding area, made me lazy to get out of bed.

(26) *Orang-orang yang dekat di hati saya, satu persatu mulai beranjak ke* ***peraduan****.*
The people I love, one by one began to move to go to ***bed***.

The common words for 'bed' in Bahasa Indonesia is *tempat tidur* or *ranjang*. However, in sentences (24), (25) and (26) which contexts is not about the royal family, the word *peraduan* is used instead of *tempat tidur* or *ranjang*. This usage shows that the meaning of *peraduan*

has been generalized. Since there is not much longer king or royal family in Indonesia, the word has become usable for everyone.

The third type is the use of the word in figurative meaning. In this type of usage, the word *peraduan* mainly collocates with *matahari* (sun) like in sentence (27), (28) and (29); and *sang surya* which also means 'sun' in (30). In those sentences, the sun is depicted as if it goes to bed to rest so the day turns into night, or gets out of the bed and starts to shine.

(27) *Matahari beranjak ke **peraduan** dan malam mulai menggeliat ke atas bumi.*
The sun goes down to its ***resting place*** and the night begins to climb the earth.

(28) *Ketika matahari telah kembali ke **peraduan**, malam pun tiba.*
When the sun has gone to ***bed***, the night has come.

(29) *Matahari sudah beranjak ke **peraduan**, tetapi langit biru masih tersisa.*
The sun has gone to ***bed***, but there is still some blue sky.

(30) *Salah satunya adalah untuk melihat secara langsung, Sang Surya keluar dari **peraduan** di ufuk timur.*
One of the reason is to see directly the sun out of its ***bed*** in the eastern horizon.

Those different types of usage of *peraduan* found in Leipzig Corpora show that the word has changed in meaning through generalization and metaphor. However, unlike the word *bersiram*, the change that the word *peraduan* experienced only happened in semantic aspect. The other linguistic aspects of the word are not affected.

III. Discussions and Conclusions

Etymological information of the word *bersiram* and *peraduan* obtained from the analysis results are as follows.

Entry*: bersiram*
Initial meaning: to take a bath (intransitive), used for the royal family
Additional meaning in the 21st century: 1. bathe (transitive, figurative meaning)
2. cover (transitive, figurative meaning)

Entry: *peraduan*
Initial meaning: bed or bedroom, used for the royal family
Additional meaning in the 21st century: 1. bed or bedroom, for general
2. resting place (figurative meaning)

The presentation of the etymological information in the dictionary can also be made in the narrative form, so the reader could get a clearer picture of the changes.

The results prove that diachronic corpora are a useful tool in the investigation of etymological information, especially to find changes in meaning. The corpora that are set chronologically can also tell the approximate time of change. Although the precise year of change remains unknown, it can at least tell in which era the change happen.

However, the activity of collecting etymological information from diachronic corpora can only be done to the lexicon in written text. Information about the usage of the words in spoken forms, whether they are used in the same register with the same meaning or not, is not known. But to compile a diachronic spoken corpus from the past is impossible. Nevertheless, it does not diminish the effectiveness of diachronic corpora as a tool in collecting etymological information.

References

Allan, K. and Robinson, J.A. (Eds.). (2012). *Current Methods in Historical Semantics*. Berlin/Boston: Walter de Gruyter GmbH & Co

Durkin, Philip. (2009). *The Oxford guide to etymology*. New York: Oxford University Press

Jones, R. (1978). *Arabic loan-words in Indonesian: a check-list of words of Arabic and Persian origin in Bahasa Indonesia and traditional Malay, in the reformed spelling*. London: School of Oriental and African Studies, University of London.

Kamus Dewan Edisi Keempat. (2015). Kuala Lumpur: Dewan Bahasa dan Pustaka

Kamus Besar Bahasa Indonesia Edisi V. (2016). retrieved from https://kbbi.kemdikbud.go.id/

Kridalaksana, H. (2001). "Arah Pengembangan Kajian Etimologi Indonesia". *Kata*. April, 2001.

Leipzig Corpora, retrieved from http://corpora.uni-leipzig.de/en?corpusId=ind_mixed_2013

Liberman, A. (2009). *Word origins and how we know them: Etymology for everyone*. New York: Oxford University Press.

Malay Concordance Project, retrieved from http://mcp.anu.edu.au/

McEnery, T. and Hardie, A. (2012). *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.

Proudfoot, I. (1991). Concordances and classical Malay. *Bijdragen tot de Taal-, Land- en Volkenkunde* 147 (1991), no: 1, Leiden, 74-95.

The Turkish Lexicography Corpus (TLC): An Overview

Prof. Dr. Erdoğan BOZ; Ferdi BOZKURT, Ph.D.; Fatih DOĞRU, Ph.D.
Eskişehir Osmangazi University, Anadolu University, Eskişehir Osmangazi University
erdoganboz@ogu.edu.tr, ferdib@anadolu.edu.tr, fdogru@ogu.edu.tr

Abstract

There is a lack in the field of lexicography in terms of terminology use because there is not a lexicography terminology created and made available for researchers in Turkey although there is an increase in lexicography studies. Therefore, in order to fulfil this need and create a Turkish Lexicography terminology, Eskisehir Osmangazi University Center for Lexicography has decided to start a project.

This study aims to create a specialised corpus including Turkish lexicography studies and to set forth Turkish Lexicography terminology by using this corpus. To create a specialised corpus will help the researchers in deciding what terminology they can use in their studies and it can also help to standardize the terminology use.

There is not a platform in which the researchers can discuss and see the previous terminology use in the studies. Terminology choice is mainly made intuitively and it usually depends on small academic group discussions. A corpus can ease the terminology use of the researchers in the field.

Based on these necessities, a corpus for Turkish lexicography was created and is accessible to the researchers on a website, www.tsd.ogu.edu.tr . The stages of the study were as follows: synchronization and desynchronization of the corpus, determining the corpus content, digitizing the sources, external tagging of the texts, text type tagging, and lemmatizing. There is also a platform on the website via which the researchers can send new terms they have encountered in related studies.

Keywords: Lexicography, terminology, term, corpus, TLC

1. Introduction

In recent years, the database creation studies in the world have been done via electronic corpora. Sinclair defines corpus as a collection of pieces of language text in electronic form (Sinclair, 2005: 16). Corpora are data sources prepared for different purposes and for making language, vocabulary generalizations in different forms. Specialised corpus is built for a particular research project, subject areas, domains, topics etc. (Baker and Hardie, 2006: 147). There are many specialised corpora in the world besides general corpora. For example,

- Helsinki Corpus of English Texts (The first specialised electronic diachronic corpus of English) (Kytö & Voutilainen, 1995),
- the Aarhus Corpus of Contract Law (Anderson, 2006: 83),
- Air Traffic Control Speech corpus (Hofbauer et. al, 2008),
- Lampeter Corpus of Early Modern English Tracts (Siemund & Claridge, 1997),
- USE – Uppsala Student English corpus (Axelsson, 1999),
- Guangzhou Petroleum English Corpus (Baker et. al, 2006).

2. Building of the TLC

The aim of the TLC is to provide a collection which includes master theses, doctoral dissertations, published presentations, news, books, articles, and reviews about the field of Turkish lexicography. It will be possible to obtain a word list of Turkish lexicographic terms in this specialised corpus.

There are some phases of building of the TLC. The first phase was determination of the content. The content of the TLC is master theses, doctoral dissertations, published presentations, news, books, articles, and reviews about the field of Turkish lexicography. As the boundaries of field of lexicography are not clear cut, a criterion was defined to select the text for the TLC. The criterion was to filter the texts whether they had some specific keywords; "sözlük" (dictionary), "lügat" (dictionary, an old usage), "sözlükbilim" (lexicography), "sözlük bilim" (lexicography), "sözlükbilimi" (lexicography), "sözlük bilimi" (lexicography), "sözlükçülük" (synonym with lexicography), "leksikografi" (lexicography). The texts published between 1932 and 2016 were used. The year of 1932 is the year when the Turkish Language Institute (Türk Dil Kurumu) was founded and the 2016 is the year when the project began.

Text Type	***Number of Texts***
Master theses	39
Doctoral dissertations	12
Published presentations	310
News	21
Books	3
Articles	468
Reviews	150
Total	**1003**

Table 1: Text types included in the corpus database

Some of the texts were not digitally available. They were scanned and converted to PDF format. Afterwards the PDF texts were converted to OCR format. In the converting phase there were some missing characters or spelling errors in the texts. They were corrected by the researchers in the correction phase in three months. After the correction, the texts were

uploaded to database of the corpus in one month. Metadata and layers of texts were determined, and the texts were classified based on text genres, and publication year, author. As Turkish language is an agglutinative language, the next phase was lemmatizing. In the lemmatizing phase, the lemmas and the suffixes were determined. In the tagging phase, some lemmas related to the field of lexicography were selected from the sample sentences by means of "term extraction tab" and researchers decided on whether these lemmas can be terms or not. This procedure has shown that there are both single-word and multi-word terms in the field of Turkish lexicography. N-gram tool of the software was utilized to detect the multi-word terms occurring in the TLC.

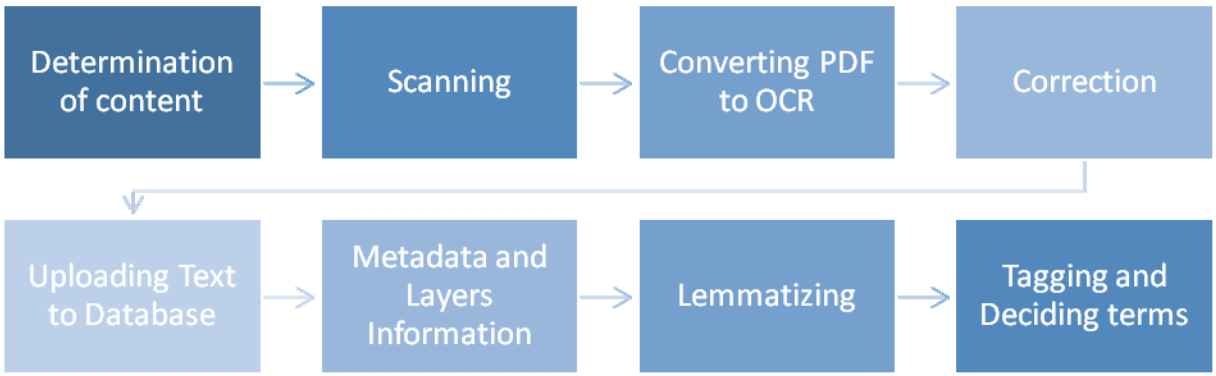


Figure 1: TLC building phases

3. Overview of the TLC

"Turkish Lexicography Corpus" was built and made available for the users as an outcome of a project, funded by Eskişehir Osmangazi University with the code of 2016-019056, titled "A Corpus-based Research on Terminology of Turkish Lexicography". The website of the TLC is available on www.tsd.ogu.edu.tr.

3.1. Content of the TLC

There are 1003 texts, 42.831 sentences, 703.986 orthographic words and 86.368 types in the TLC. Totally 1.616 lexicographic terms were determined in the TLC by the project researchers.

Total number of texts	Total number of sentences	Total number of words	Total number of types
1003	42.831	703.986	86.368

Figure 2: Content

3.2. Registration

In order to make a search in the TLC website, it is obligatory to register.

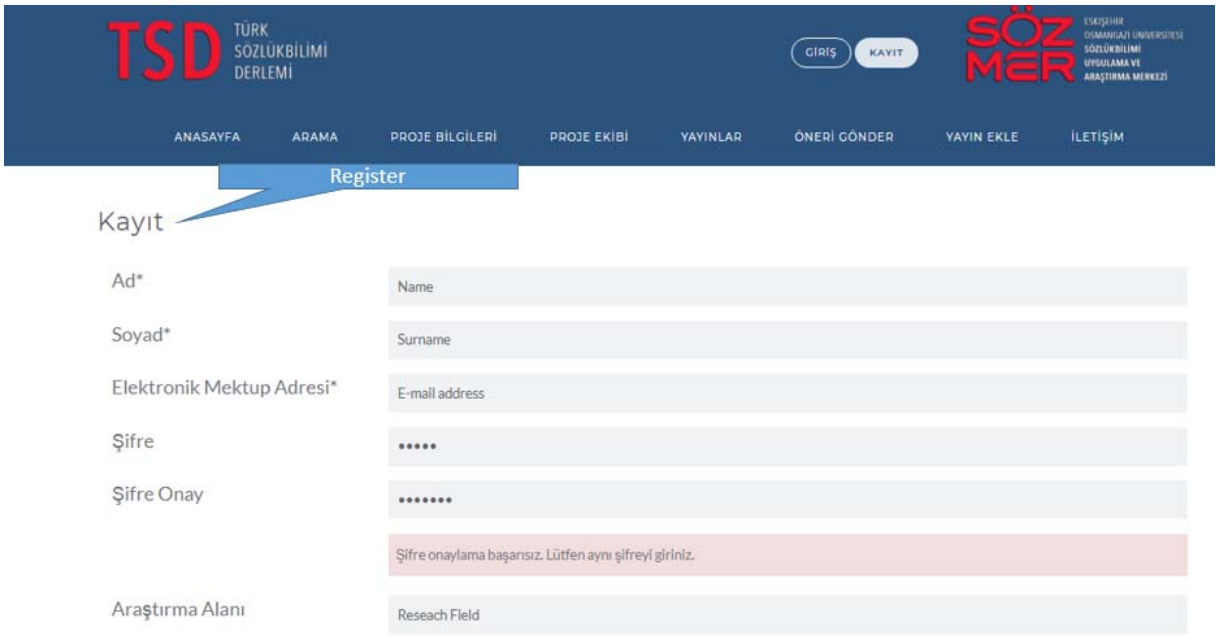


Figure 3: Registration menu

3.3. Log in

Once your registration is approved, you can log in the TLC by using your username and password in the "Log in" menu. (An e-mail is not sent to the users for registration approval.)

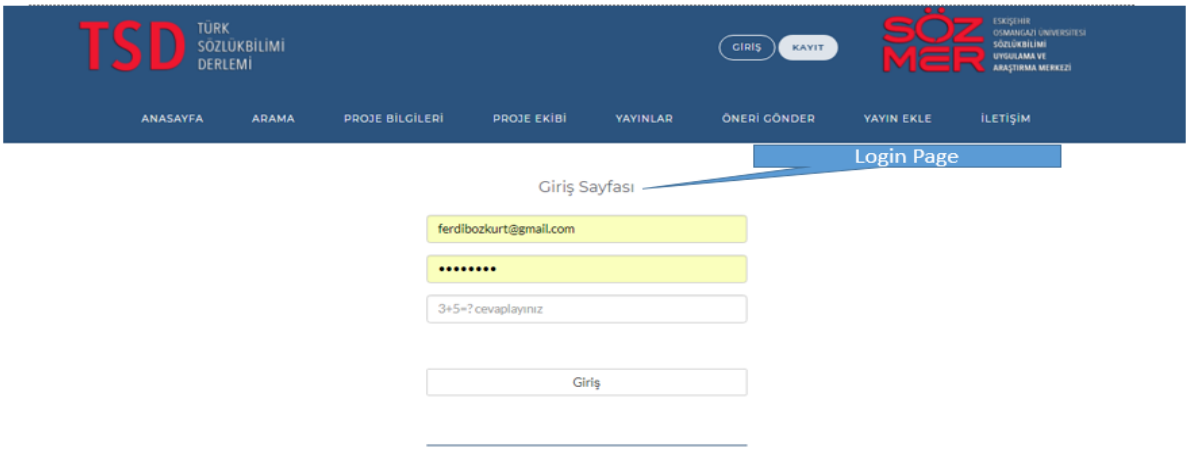


Figure 4: Login menu

3.4. Query

3.4.1. Key words in context tool (KWIC Tool)

KWIC can display basic information about the frequency of the search term and its distribution in texts. The searched term 'buldurucu' occurs 3 times in the TLC with the frequency of 3 in 1 text out of 1003 texts.

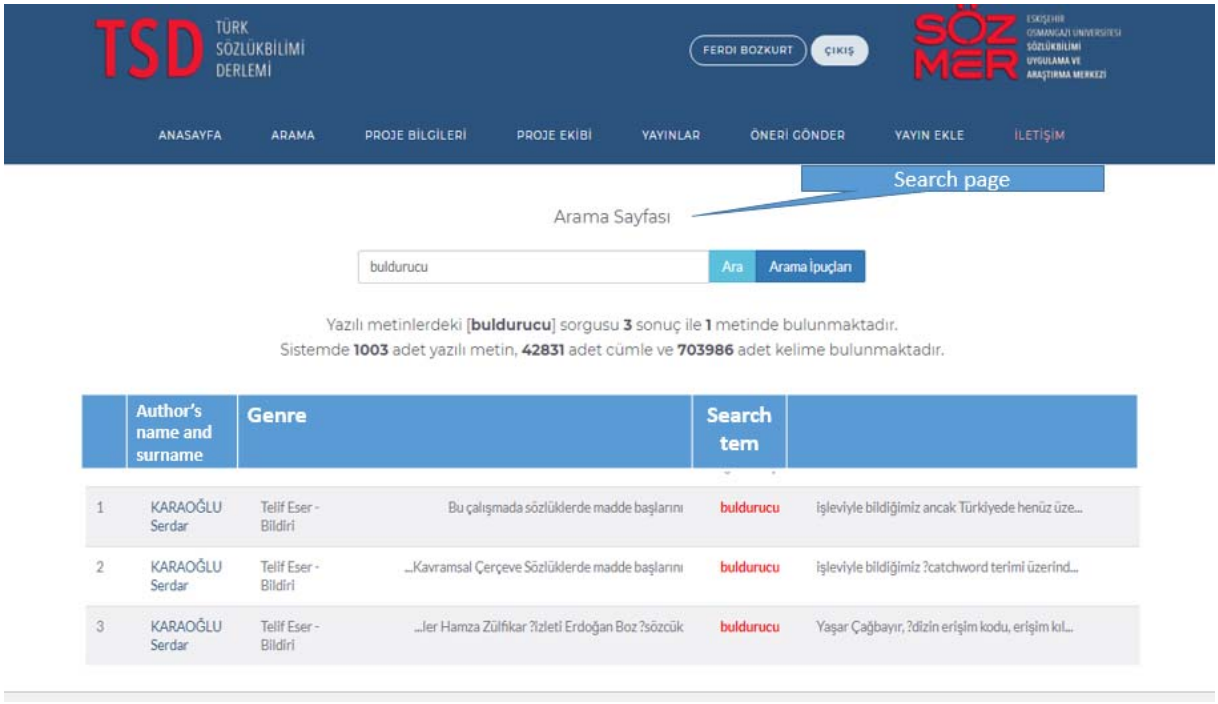


Figure 5: Key words in context tool

3.4.2. Search Tips

"Search Tips" can be used for detailed search.

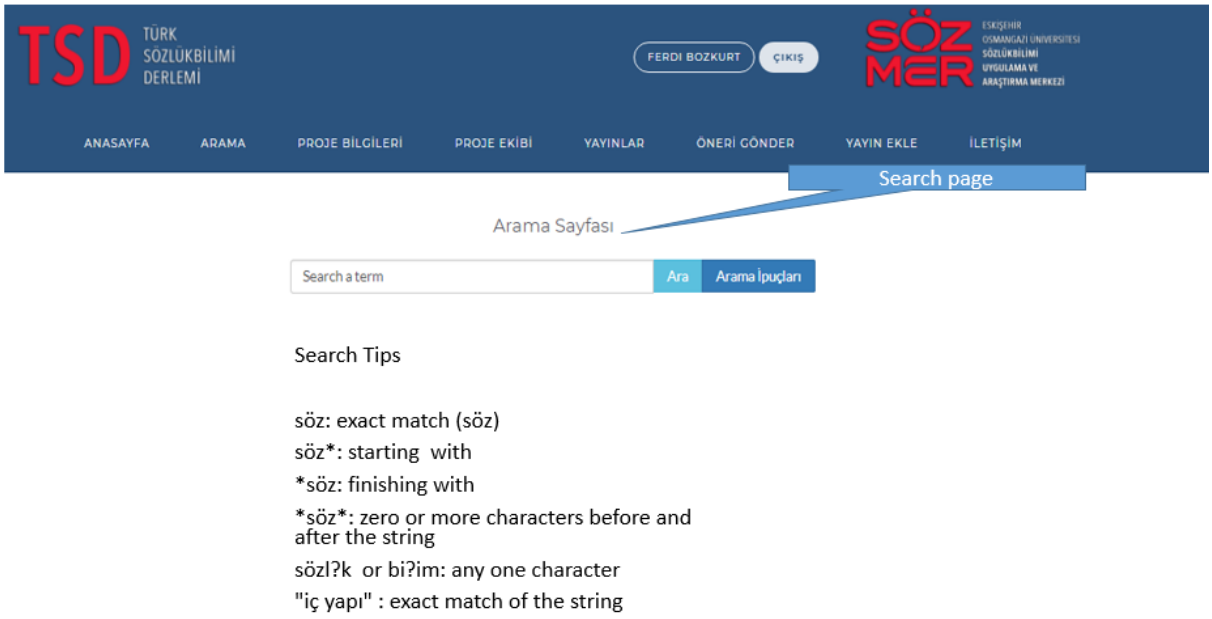
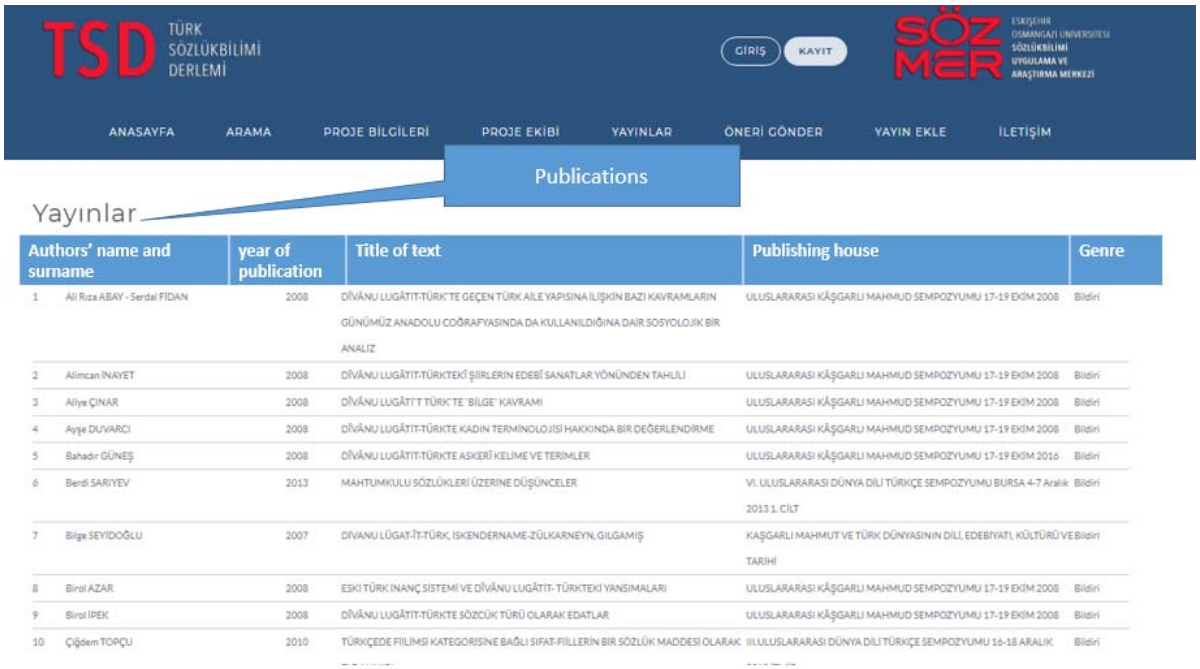


Figure 6: Search Tips

3.5. Publications

In the "Publications" menu, you can see the publications that corpus included with the information of authors' name, year of publication, the name of text, publishing house and text genres.



	Authors' name and surname	year of publication	Title of text	Publishing house	Genre
1	Ali Rıza ABAY - Serdal FİDAN	2008	DİVÂNU LUGÂTİT-TÜRK'TE GEÇEN TÜRK AİLE YAPISINA İLİŞKİN BAZI KAVRAMLARIN GÜNÜMÜZ ANADOLU COĞRAFYASINDA DA KULLANILDIĞINA DAİR SOSYOLOJİK BİR ANALİZ	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
2	Alimcan İNAYET	2008	DİVÂNU LUGÂTİT-TÜRKTEKİ ŞİİRLERİN EDEBİ SANATLAR YÖNÜNDEN TAHLİLİ	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
3	Aliye ÇINAR	2008	DİVÂNU LUGÂTİT TÜRK'TE 'BİLGE' KAVRAMI	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
4	Ayşe DUVARCI	2008	DİVÂNU LUGÂTİT-TÜRKTE KADIN TERMİNOLOJİSİ HAKKINDA BİR DEĞERLENDİRME	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
5	Bahadır GÜNEŞ	2008	DİVÂNU LUGÂTİT-TÜRKTE ASKERİ KELİME VE TERİMLER	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2016	Bildiri
6	Berdi SARIYEV	2013	MAHTUMKULU SÖZLÜKLERİ ÜZERİNE DÜŞÜNCELER	VI. ULUSLARARASI DÜNYA DİLİ TÜRKÇE SEMPOZYUMU BURSA 4-7 Aralık 2013 1. CİLT	Bildiri
7	Bilge SEYİDOĞLU	2007	DİVANU LÜGAT-İT-TÜRK, İSKENDERNAME-ZÜLKARNEYN, GILGAMIŞ	KAŞGARLI MAHMUT VE TÜRK DÜNYASININ DİLİ, EDEBİYATI, KÜLTÜRÜ VE TARİHİ	Bildiri
8	Birol AZAR	2008	ESKİ TÜRK İNANÇ SİSTEMİ VE DİVÂNU LUGÂTİT- TÜRKTEKİ YANSIMALARI	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
9	Birol İPEK	2008	DİVÂNU LUGÂTİT-TÜRKTE SÖZCÜK TÜRÜ OLARAK EDATLAR	ULUSLARARASI KAŞGARLI MAHMUD SEMPOZYUMU 17-19 EKİM 2008	Bildiri
10	Çiğdem TOPÇU	2010	TÜRKÇEDE FİİLİMSİ KATEGORİSİNE BAĞLI SIFAT-FİİLLERİN BİR SÖZLÜK MADDESİ OLARAK	III.ULUSLARARASI DÜNYA DİLİ TÜRKÇE SEMPOZYUMU 16-18 ARALIK	Bildiri

Figure 7: Publications

3.6. Contribution from users

The users of TLC will be able to suggest terms or present their opinions about an existing term which is included in the TLC through the "Contribution" menu. They will also be able to contribute terms that they encounter in texts related to lexicography by filling into the "user terminology form" if they are not in the TLC.

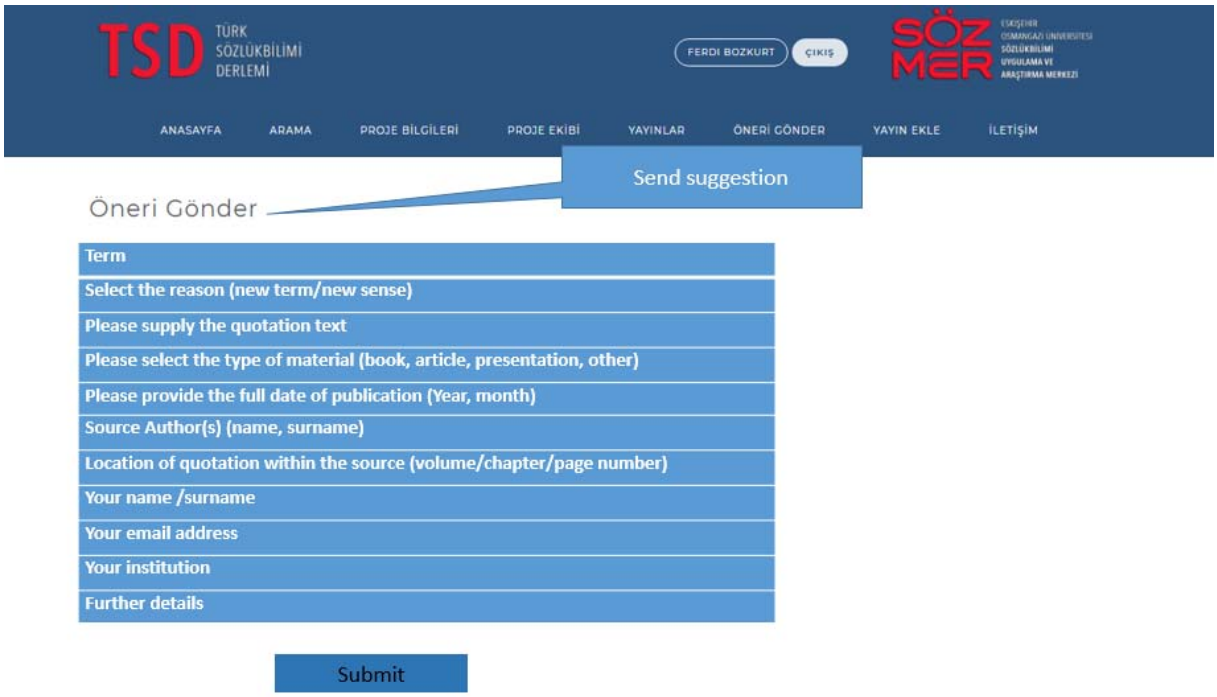


Figure 8: Contribution sending menu

3.7. Add a Publication

By using the "Add a Publication" menu, users can add the information of the publications that should be found in the TLC database.

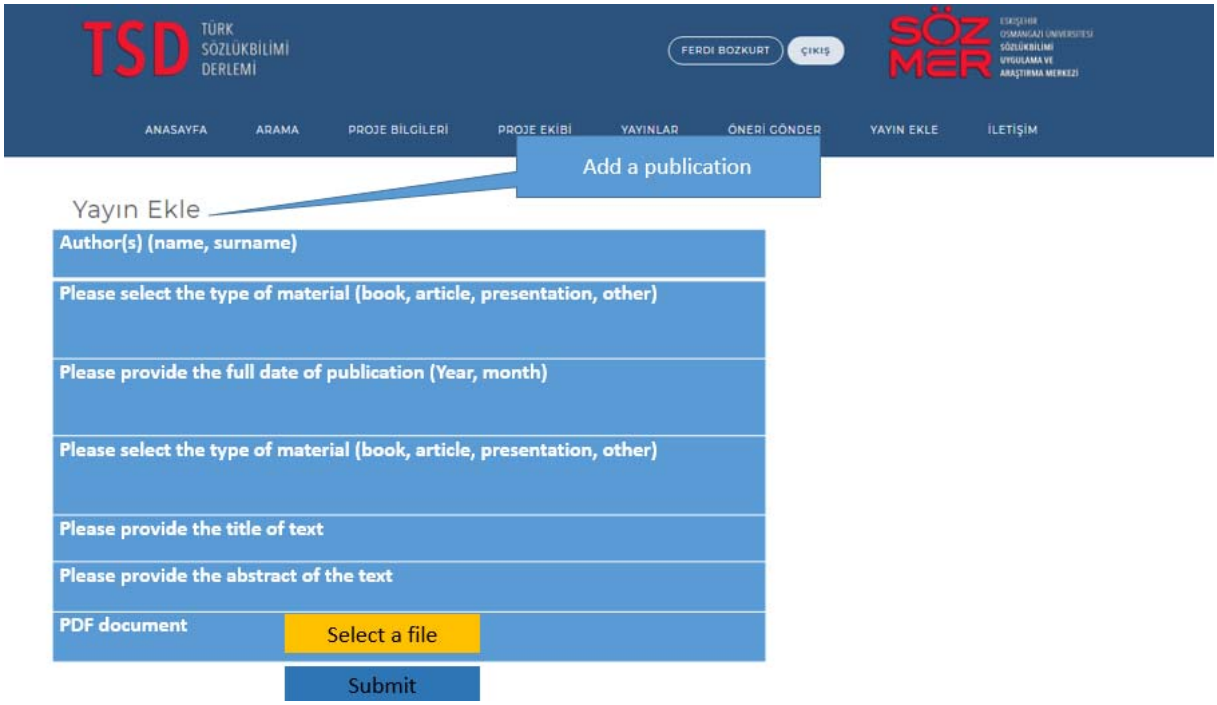


Figure 9: Add publication menu

3.8. Project Team

Prof. Dr. Erdoğan Boz is the project coordinator. Ferdi Bozkurt, Ph.D. and Fatih Doğru, Ph.D. are the researchers of Project. Şerife Sazak is the scholarship student of the project and İbrahim Yapıcı is the software specialist of the project.

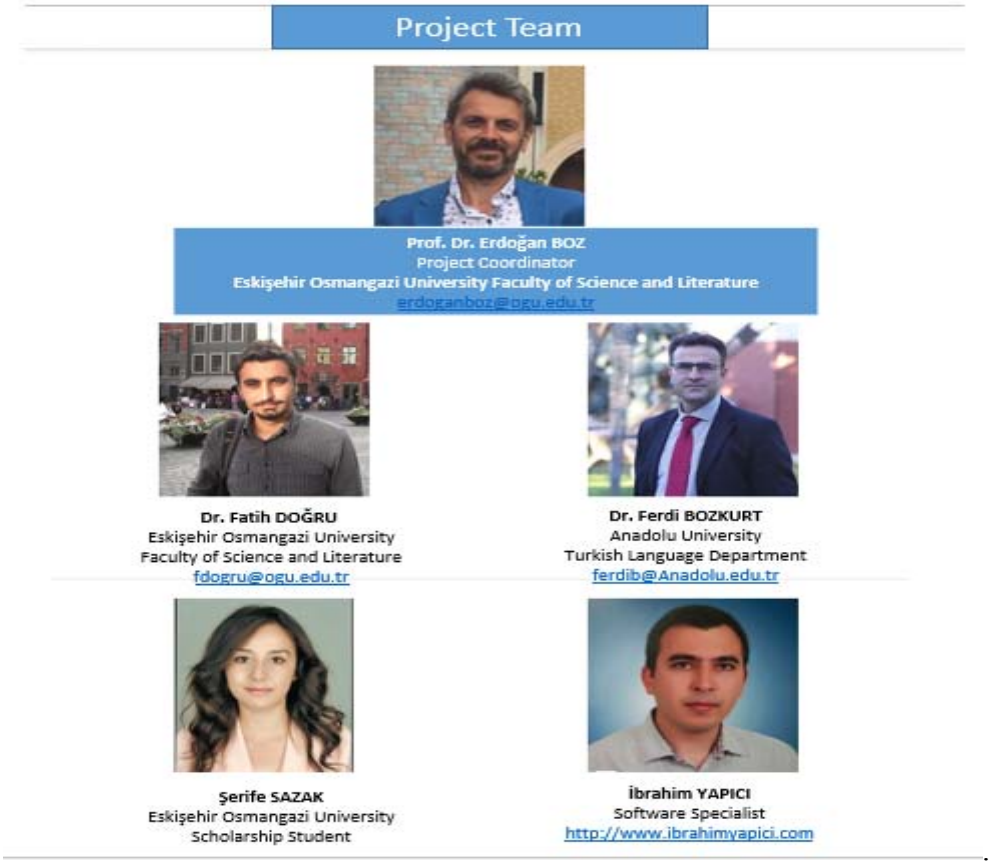


Figure 10: Project team

4. Conclusion

The TLC has been introduced in this study. The TLC is a specialised corpus including Turkish lexicography studies and it presents Turkish Lexicography terminology to researchers.

Building a specialised corpus will help the researchers in deciding what terminology they can use in their studies and it can also help to standardize the terminology use. At first, brief information on the stages of the building of the TLC has been given in this study. The stages of the project were as follows: determining diachronic boundary of the corpus, determination of the content, scanning, converting PDF to OCR, correction, uploading text to database, metadata and layers information, lemmatizing, tagging and deciding terms.

Information has been given about the corpus that has been made available to researchers at the http://www.tsd.ogu.edu.tr web address. In this study, information about the TLC has been presented under the titles of content, registration, log in, query, publications, contribution sending, add publication and project team.

References

Anderson, W. J. (2006).*The Phraseology of Administrative French: A corpus-based Study*. No. 57, Rodopi.

Axelsson, M. W. (1999). Project USE (Uppsala Student English). *ASLA Information* 25:2, 25-26.

Boz, E., Bozkurt, F. and Doğru F. (2017). "Türk Sözlükbilimi Terminolojisi Üzerine Derlem Tabanlı Bir Araştırma", *III. Uluslararası Sözlükbilimi Sempozyumu Bildiri Kitabı*. Eskişehir: Eskişehir Osmangazi University Press.

Hofbauer, K., Petrik, S. and Hering, H. (2008). The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech. In *LREC*

Kytö, M. and Voutilainen, A. (1995). Applying the Constraint Grammar parser of English to the Helsinki corpus. *ICAME Journal*, *19*, 23-48.

Siemund, R. and Claridge, C. (1997). The Lampeter Corpus of Early Modern English Tracts. *ICAME JOURNAL*, *21*, 61-70.

Sinclair, J. M. (2005). "Corpus and Text-basic Principles", *Developing Linguistic Corpora: A Guide to Good Practice*, Editör *Martin Wynne*, Oxbow Books, Oxford, ss. 1-16.

Polyonymy in terminology of Turkish lexicography

Fatih DOĞRU, Ph.D.
Eskişehir Osmangazi University
fdogru@ogu.edu.tr

Abstract

The definition of polyonymy is the use of multiple nouns for the same concept. It can be said that it is the same as the synonymy when it is addressed the semantics perspective, the difference between polyonymy and synonymy is related to the subject approach. Terminology and semantics address this concept differently.

Standardization is extremely important in terms used in studies in specific fields such as lexicography. Words can be taken from the general language or another field on terminologization. The polyonymy formed in this way is reasonable. However, having multiple signifiers of the same concept in one field can cause various difficulties. Nevertheless, a concept in a field of science is being used with multiple signifiers for various reasons. The difference between these signifiers may be due to spelling, loanword usage, usage of direct translation of loanword, usage of different suffixes, ideal of nativization and the preference of the author.

In order to reveal the polyonymy on a science field, a specialised corpus makes it easy to detect the examples. In this study, Turkish Lexicography Corpus (TLC) which was built by Eskişehir Osmangazi University Center for Lexicography will be utilized in order to reveal the polyonymy in the field of Turkish lexicography. TLC contains articles, conference papers, books, bulletins and postgraduate theses written in Turkish in the field of Turkish lexicography between 1932 and 2016. This corpus can monitorize the terms of Turkish lexicography and it contains different text types. Making a list of polyonym words and to reveal the causes of polyonymy is important for homogeneity in communication on Turkish lexicography field. In addition, the presentation of identified polyonym terms may facilitate the choice of term in future studies of researchers.

Keywords: Polyonymy, lexicography, terminology, corpus

1. Introduction

Polyonymy is the use of multiple nouns for the same concept. It is defined in the OED "Each of a number of different words having the same meaning; a synonym." Synonymy represents a deviation from the canonical one form - one meaning relation (Murphy, 2003: 162). Even though theoretically a concept is expressed by a single designation, in reality there are alternative designations for a single concept and the designation of two different concepts can coincide, even within the same special field. Generally, two units designating the same concept are synonyms (Cabré, 1999: 109). They can also be called polyonyms. There are at least two units designating the same concept in polyonymy and in some cases the units number may increase.

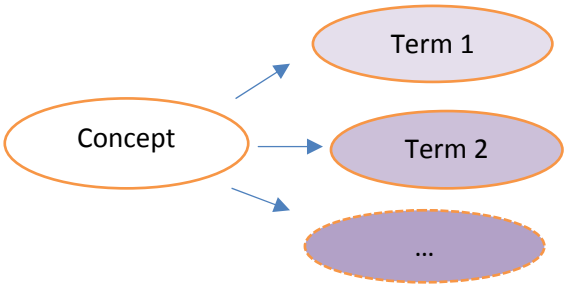


Figure 1: Polyonymy in a concept

While polyonymy can be reasonable in general communication, the situation in special communication is different. Using more than one term designating the same concept in special communication may cause ambiguity. "Communication without ambiguity would require each designation to correspond to a single concept and each concept could only be designated by a single term. This is clearly not the case for general language, in which words are usually polysemous and meanings can be expressed by several alternatives that are synonymous to one another" (Cabré, 1999: 194).

Polyonymy in a special field may cause some difficulties in communication between researchers. Terminological standardization is required in terms of a special field to prevent these difficulties. "Ambiguous terminology based on polysemy, synonymy -in other words polyonymy- and homonymy obviously presents obstacles for communication among specialists and inevitably frustrates efforts to order thought. For this reason as early as the 19th century, scientists and, at the beginning of the 20th century, technicians, felt it was necessary to regularize terminology in their respective areas and thus became directly involved in the standardization process" (Cabré, 1999: 194). Disambiguity is vital for terminologies of sciences, however, absolute synonyms still appear, for various reasons (Vogel, 2006: 40).

Some researchers claim that polyonymy is functional. "Polyonymy or (near-) synonymy exists because the mechanisms for naming can trigger several possible lexicalizations. Slightly different perspectives result in polyonymy or near-synonyms. The univocity ideal of traditional terminology consists of trying to eliminate some of the polyonyms or near-synonyms and indicating a preferred term. The underlying idea is that to have several terms for the same concept/category is undesirable as it implies an impediment for unambiguous communication" (Temmerman, 2000: 150). Temmerman (2000: 150) says that the functional aspect of synonymy in a discourse community that is overlooked. Temmerman claims that polyonymy is functional and it has some functional advantages.

Sager (1990: 60) sees "terminologisation" as a process in time. He sees the evolution of concepts as accompanied by stages of naming. At the end of this process, there can also be polyonymy in terminology. Sager (1990: 89) presented a list which was an example of a highly idealised requirement which, however, can only be realised in a strictly controlled

environment. He said that in the mentioned list there should be no synonyms or polyonyms whether absolute, relative or apparent.

Nevertheless there are many polyonym terms in the terminology of lexicography. In this study, polyonymy in the terminology of Turkish lexicography will be examined. Turkish Lexicography Corpus (TLC/TSD: http://tsd.ogu.edu.tr) was built by Eskişehir Osmangazi University Center for Lexicography (SÖZMER) in order to reveal terminology of Turkish lexicography. TLC contains articles, conference papers, books, bulletins and postgraduate theses written in Turkish in the field of Turkish lexicography between 1932 and 2016. It contains the texts which include the keywords related to lexicography as "sözlük" (dictionary), "lügat" (dictionary), "sözlükbilim" (lexicography), "sözlük bilim" (lexicography), "sözlükbilimi" (lexicography), "sözlük bilimi" (lexicography), "sözlükçülük" (lexicography), "leksikografi" (lexicography). This corpus can monitorize the terms of Turkish lexicography and it contains different text types. There are 1003 texts in the corpus. It comprises 42.831 sentences, 703.986 orthographic words, and 86.368 types. 1616 lexicographic terms are listed through the corpus. Polyonymy in these 1616 terms is discussed in this study. There are some reason of polyonymy in Turkish lexicography.

2. Polyonyms in the TLC

There are some reasons of polyonymy in Turkish lexicography. These reasons are concurrently related to terminologization and term formation. Their functionality is not discussed in this study. The determined examples of polyonymy are listed and the reasons for the formation are discussed.

2.1. Orthographical differences

Variety of orthographical spelling differences are a common reason of polyonymy. These differences occur when a term has not got a common use orthography in the literature. Especially it is observed in compound words. Compounding which is the combination of existing words into new ones is a common method of designation of new concepts (Sager, 1990: 72). Spelling differences of multi-word terms and phonetic changes reveal polyonymy in terminology of Turkish lexicography. Examples are listed in Table 1 and those related to orthographical differences are written in italics.

Term (Text(s) / Frequency)
ad bilimi (2/5), *ad bilim* (2/2), *adbilimi* (1/5)
ağız (47/102), *diyalekt* (13/30), *dialekt* (3/3)
alt anlam (2/2), *altanlam* (1/1)
alt anlamlılık (3/15), *altanlamlılık* (2/5)
alt madde (6/16), *altmadde* (1/3)
altmaddebaşı (2/3), *alt madde başı* (1/8), *alt maddebaşı* (1/3)
semantik (27/65), *anlambilim* (19/32), *anlambilimi* (9/19), *anlam bilimi* (7/10), *anlam bilim* (2/2)
anlambirim (5/5), *anlam birimi* (3/3)
anlambirimcik (6/39), *anlam birimcik* (2/11)

anlamsal (42/131, semantik (27/65), *anlambilimsel* (12/60), *anlam bilimsel* (4/70)
art zamanlı (9/24), *artzamanlı* (8/9), diyakronik (2/2)
bilgisayar destekli sözlük bilimi (2/10), *bilgisayarlı sözlük bilimi* (2/4), *bilgisayar destekli sözlükbilim* (1/2), bilgisayar destekli sözlükçülük (1/1), *bilgisayarlı sözlükbilimi* (1/1), bilgisayar destekli sözlük yazımı (1/1)
bütüncül yapı (5/24) *dış yapı* (4/5), *makro yapı* (3/3), büyük yapı (1/2), *makroyapı* (1/1), *dışyapı* (1/1)
çekirdek sözvarlığı (2/2), çekirdek sözcük (1/3), çekirdek kelime (1/1), çekirdek söz (1/1), *çekirdek söz varlığı* (1/1)
çevrimiçi sözlük (4/14), *çevrim içi sözlük* (3/6), online sözlük (3/6)
çevrim içi sürüm (1/4), *çevrimiçi sürüm* (1/1)
çok anlamlı (20/48), *çokanlamlı* (8/43), çok manalı (1/1)
çok anlamlılık (5/14), *çokanlamlılık* (5/13)
çok dilli sözlük (15/30), *çokdilli sözlük* (1/1)
derlem dilbilim (3/10), *derlem dil bilimi* (1/2), *derlemdilbilim* (1/1), bütünce dil bilimi (1/1)
derlem tabanlı (8/18), *derlem temelli* (3/3), *derlemtemelli* (2/5), *derlemtabanlı* (1/3), (corpusbased)
dış yapı (4/5), *dışyapı* (1/1), bütüncül yapı (5/24
diyalektoloji (9/21), *dialektoloji* (2/3), ağız bilimi (1/1)
dil bilgisel eşdizimlilik (1/1), *dilbilgisel eşdizimlilik* (1/1)
dilbilgisel anlam (2/6), *dil bilgisel anlam* (1/3)
dilbilgisel bilgi (3/20), *dil bilgisel bilgi* (2/4)
dil bilgisi terimleri sözlüğü (4/5), *dilbilgisi terimleri sözlüğü* (2/8)
dilbilgisel (26/92), *dil bilgisel* (11/45), gramatikal (6/18)
düzanlam (2/6), *düz anlam* (2/3)
ek (223/2082), *biçimbirim* (18/119), *biçimbirimi* (5/14), morfem (5/13), *biçim birimi* (2/2)
enantiosemiye (1/25), *enantiosemiya* (1/5)
eşadlı (8/8), *eş adlı* (2/5)
eşadlılık (3/4), *eş adlılık* (1/4)
eşanlam (6/20), *eş anlam* (2/4)
eş anlamlı (24/58), *eşanlamlı* (18/25), anlamdaş (8/8), sinonim (5/10), müteradif (3/3)

eş anlamlılık (5/7), *eşanlamlılık* (2/4)
eşdizimlilik (7/85), *eş dizimlilik* (1/3)
eşdizimsel (5/157), *eş dizimsel* (2/14)
eş sesli (6/18), *eşsesli* (5/5), homofon (1/1)
eş seslilik (2/18), sesteşlik (1/6), eşseslilik (1/2)
eşyazımlı (4/6), *eş yazımlı* (1/1), homograf (1/1)
eş zamanlı (15/29), *eşzamanlı* (10/12), senkronik (1/1)
fonetik (43/84), *ses bilgisi* (14/19), *sesbilgisi* (4/4)
gramer (69/252), *dilbilgisi* (43/152), *dil bilgisi* (41/141)
iç yapı (2/3), *içyapı* (3/4)
iki dilli sözlük (29/121), *ikidilli sözlük* (4/7)
köken bilgisi sözlüğü (10/13), *kökenbilgisi sözlüğü* (3/4)
leksikoloji (14/24), *sözcükbilim* (12/21), *sözcük bilimi* (4/4), *sözcükbilimi* (3/4), *sözcük bilim* (2/2)
madde (113/918), *madde başı* (79/442), *sözlükbirim* (16/45), *maddebaşı* (15/120), girdi (13/28), entry (9/29), lemma (8/28), sözlük maddesi (7/13), *sözlük birim* (5/169), headword (4/13), *sözlükbirimi* (4/12),başsözcük (3/10), sözlüksel birim (4/7), lexeme (4/7), *sözlük birimi* (1/11), giri (1/9), leksem (1/3)
medioyapı (1/1), *medio yapı* (2/2)
metinlerarasılık (1/3), *metinler arasılık* (1/1)
morfoloji (7/9), *biçimbilim* (5/6), *biçim bilim* (1/1), *biçim bilimi* (1/1), *biçimbilimi* (1/1)
offline sözlük (2/3), *çevrim dışı sözlük* (1/2), *çevrimdışı sözlük* (1/1)
parçabütün ilişkisi (4/4), *parça bütün ilişkisi* (2/9), parçabütün ilgisi (1/1)
parçacıl yapı (4/10), *mikro yapı* (3/3), *mikroyapı* (3/3), küçük yapı (1/3), parçayapı (1/1)
sesbilgisel (3/3), *ses bilgisel* (1/1)
sesbilimsel (4/16), fonolojik (3/9), *ses bilimsel* (2/4)
ses bilimi (3/3), fonoloji (2/2), *sesbilimi* (1/1)
sesteş (6/13), *eş sesli* (6/18), *eşsesli* (5/5)
sözcükbirim (5/20), *sözcükbirimi* (2/9), *sözcük birim* (1/1), *sözcük birimi* (1/1)
sözcük bilgisi (3/3), *sözcükbilgisi* (1/1)
sözdizimsel (18/53), *söz dizimsel* (13/108), sentaktik (5/19)

sözlükbilim (34/167), *sözlükbilimi* (30/269), *sözlük bilimi* (20/78), *sözlük bilim* (9/15), leksikografi (7/17), leksikografya (2/2), sözlükçülük (64/458), leksikoloji (7/8), lügatçilik (5/7), *sözlükbilgisi* (5/5), *sözlük bilgisi* (1/1), sözlükçülük bilimi (1/1)
sözlükbilimsel (12/103), *sözlük bilimsel* (6/31), leksikografik (4/6)
sözlükbilimci (17/41), *sözlük bilimci* (6/33), sözlükbilim uzmanı (1/1), leksikograf (1/1)
tek dilli sözlük (16/29), *tekdilli sözlük* (2/2)
temel anlam (8/19), ilk anlam (5/5), birincil anlam (2/2), *düzanlam* (2/6), *düz anlam* (2/3), yaygın anlam (2/2), asıl anlam (1/5), öz anlam (1/1), ilkel anlam (1/1)
teorik sözlükbilim (4/5), *kuramsal sözlükbilimi* (3/6), *kuramsal sözlükbilim* (2/5), kuramsal sözlükçülük (2/2), kuramsal leksikografi (1/8), teorik leksikografi (1/3), teorik sözlükbilimi (1/2), *kuramsal sözlük bilimi* (1/1), *teorik sözlük bilim* (1/1)
terminoloji (32/130), *terimbilimi* (3/16), *terimbilim* (3/15), terminografi (2/67), *terim bilimi* (2/3), *terim bilim* (1/1)
transkripsiyon (7/15), *çeviri yazı* (3/3), *çeviriyazı* (2/6), *çevriyazı* (1/2)
uygulamalı sözlükbilimi (5/12), uygulamalı sözlükbilim (3/3), uygulamalı leksikografı (1/2), uygulamalı sözlükçülük (1/1), *pratik sözlükbilim* (1/1), *pratik sözlük bilim* (1/1), pratik sözlükçülük (1/1)
üstanlam (1/1), *üst anlam* (1/1)
üstanlamlılık (1/3), *üst anlamlılık* (1/1)
yan anlam (9/36), *yananlam* (3/4), ikincil anlam (1/1)
zıt anlamlı (12/41), karşıt anlamlı (9/11), ezdad (2/10), antonim (1/2), *zıtanlamlı* (1/1)
zıt anlamlılık (6/11), karşıt anlamlılık (2/7), *zıtanlamlılık* (1/1)

Table 1: Polyonyms based on orthographical differences

2.2. Borrowing

Borrowing is a term formation method. Linguistic communities which import scientific and technological knowledge tend to prefer the use of autochthonous linguistic resources for the creation of terminology, even if for a short time there is a certain amount of direct borrowing. Loan translation is preferred to direct borrowing, but neither form of term creation is acceptable if it violates the natural word formation techniques of a linguistic community (Sager, 1990: 87). Polyonymy reveals when loanwords and native words are used together in terminology of a special field. In terminology of Turkish lexicography Turkish-origin terms and loanwords –written in original form (direct borrowings) or according to the Turkish writing rules (adapted borrowings) are used together. Some terms are compounded with a loanword and a native word such as *makroyapı* (English+Turkish). Therefore, there are many polyonymy in this field. Examples are listed in Table 2 and those related to borrowing are written in italics.

Term (Text(s) / Frequency)

sesbirim (2/2), *fonem* (1/1)
semantik (27/65), anlambilim (19/32), anlambilimi (9/19), anlam bilimi (7/10), anlam bilim (2/2)
anlam farkı (1/1), anlam *nüansı* (1/1)
arkaik kelime (3/6), *arkaik leksikon* (1/1), arkaik sözcük (1/1)
bütüncül yapı (5/24) dış yapı (4/5), *makro* yapı (3/3), büyük yapı (1/2), *makroyapı* (1/1), dışyapı (1/1)
çapraz gönderme (2/2), çapraz gönderim (1/5), çapraz *referans* (1/2)
çevrimiçi sözlük (4/14), çevrim içi sözlük (3/6), *online* sözlük (3/6)
derlem (26/234), bütünce (7/34), *korpus* (1/1)
ağız (47/102), *diyalekt* (13/30), *dialekt* (3/3)
diyalektoloji (9/21), *dialektoloji* (2/3), ağız bilimi (1/1)
ağız sözlüğü (4/5), *diyalektolojik* sözlük (2/3), *diyalekt* sözlüğü (1/1)
IPA (4/8), Uluslararası *Fonetik* Alfabe (1/2), Uluslararası Sesbilgisi Alfabesi (1/1),
ek (223/2082), biçimbirim (18/119), biçimbirimi (5/14), morfem (5/13), biçim birimi (2/2)
eş anlamlı (24/58), eşanlamlı (18/25), anlamdaş (8/8), *sinonim* (5/10), *müteradif* (3/3)
eş sesli (6/18), eşsesli (5/5), *homofon* (1/1)
eşyazımlı (4/6), eş yazımlı (1/1), *homograf* (1/1)
zıt anlamlı (12/41), karşıt anlamlı (9/11), *ezdad* (2/10), *antonim* (1/2), zıtanlamlı (1/1)
enantiosemiye (1/25), *enantiosemiya* (1/5)
fonetik (43/84), ses bilgisi (14/19), sesbilgisi (4/4)
sesbilimsel (4/16), *fonolojik* (3/9), ses bilimsel (2/4)
ses bilimi (3/3), *fonoloji* (2/2), sesbilimi (1/1)
madde (113/918), madde başı (79/442), sözlükbirim (16/45), maddebaşı (15/120), girdi (13/28), *entry* (9/29), *lemma* (8/28), sözlük maddesi (7/13), sözlük birim (5/169), *headword* (4/13), sözlükbirimi (4/12),başsözcük (3/10), sözlüksel birim (4/7), *lexeme* (4/7), sözlük birimi (1/11), giri (1/9), *leksem* (1/3),
görsel *materyal* (3/3), görsel malzeme (2/3),
gramer (69/252), dilbilgisi (43/152), dil bilgisi (41/141)
dizin (50/209), *indeks* (13/25)
dizinleme (2/4), *indeksleme* (1/1)

istem (4/207), *valenz* (2/4)
istem sözlüğü (2/19), *valenz* sözlüğü (1/1)
kavram (164/1670), *konsept* (6/24)
thesaurus (5/26), kavramlar dizini (6/12), kavram dizini (3/3)
teorik sözlükbilim (4/5), kuramsal sözlükbilimi (3/6), kuramsal sözlükbilim (2/5), kuramsal sözlükçülük (2/2), kuramsal leksikografi (1/8), *teorik leksikografi* (1/3), *teorik* sözlükbilimi (1/2), kuramsal sözlük bilimi (1/1), *teorik* sözlük bilim (1/1)
Latin alfabesi (3/3), *Latin elifbası* (1/1)
*medio*yapı (1/1), *medio* yapı (2/2)
*meta*sözlükçü (1/1), *metaleksikograf* (1/1)
metaleksikografi (3/12), metasözlükçülük (1/41), üst sözlükçülük (1/2)
modern sözlükbilim (2/3), *modern* sözlükçülük (1/1), *modern* sözlükbilimi (1/1), *modern leksikografi* (1/1), çağdaş sözlükbilim (1/1)
parçacıl yapı (4/10), *mikro* yapı (3/3), *mikro*yapı (3/3), küçük yapı (1/3), parçayapı (1/1)
sözlükçe (11/32), *glossary* (8/10), lügatçe (5/6), lugatçe (2/14)
sözlükbilim (34/167), sözlükbilimi (30/269), sözlük bilimi (20/78), sözlük bilim (9/15), *leksikografi* (7/17), *leksikografya* (2/2), sözlükçülük (64/458), *leksikoloji* (7/8), lügatçilik (5/7), sözlükbilgisi (5/5), sözlük bilgisi (1/1), sözlükçülük bilimi (1/1)
yeni sözcük (11/32), yeni kelime (11/17), yeni söz (3/4), yeni sözbirim (1/1), *neoloji* (1/1), *neology* (1/1), *neologie* (1/1)
neolojizm (6/22), *neologism* (4/20), *neologizm* (1/1), yenicilik (1/3)
offline sözlük (2/3), çevrim dışı sözlük (1/2), çevrimdışı sözlük (1/1)
çevrimiçi sözlük (4/14), *online* sözlük (1/2)
adbilimsel sözlük (2/2), adbilim sözlüğü (1/1), *onomastik* sözlük (1/1)
uygulamalı sözlükbilimi (5/12), uygulamalı sözlükbilim (3/3), uygulamalı *leksikografı* (1/2), uygulamalı sözlükçülük (1/1), *pratik* sözlükbilim (1/1), *pratik* sözlük bilim (1/1), *pratik* sözlükçülük (1/1)
sözcüksel (14/47), *leksik* (12/49), *leksikal* (6/11)
sözlüksel (39/286), *leksikal* (6/11), sözlüklük (2/3)
sözcükselleşme (1/1), *leksikalizasyon* (1/1), sözcükleştirme (1/1), sözcükbirimleştirme (1/1)
leksikoloji (14/24), sözcükbilim (12/21), sözcük bilimi (4/4), sözcükbilimi (3/4), sözcük bilim (2/2)

leksikolojik (3/5), sözcük bilimsel (1/1)
sözlükbilimsel (12/103), sözlük bilimsel (6/31), *leksikografik* (4/6)
leksikolog (2/2), sözcükbilimci (1/1)
sözlükbilimci (17/41), sözlük bilimci (6/33), sözlükbilim uzmanı (1/1), *leksikograf* (1/1)
leksikon (5/43), dağarcık (1/2)
morfoloji (7/9), biçimbilim (5/6), biçim bilim (1/1), biçim bilimi (1/1), biçimbilimi (1/1)
dilbilgisel (26/92), dil bilgisel (11/45), *gramatikal* (6/18)
anlamsal (42/131, *semantik* (27/65), anlambilimsel (12/60), anlam bilimsel (4/70)
sözdizimsel (18/53), söz dizimsel (13/108), *sentaktik* (5/19)
sözlükselleşme (5/95), sözlükleşme (1/1), *leksikalizasyon* (1/1)
standartlaşma (4/5), ölçünleşme (1/1)
terimbilimci (3/8), *terminograf* (1/1)
terminoloji (32/130), terimbilimi (3/16), terimbilim (3/15), *terminografi* (2/67), terim bilimi (2/3), terim bilim (1/1)
transkripsiyon (7/15), çeviri yazı (3/3), çeviriyazı (2/6), çevriyazı (1/2)
değişken (20/35), değişke (13/16), *varyant* (9/20)
değişkenlik (7/7), varyasyon (2/3)
yönetici paneli (1/11), *admin paneli* (1/5), yönetim paneli (1/1)

Table 2: Polyonyms based on borrowing

2.3. Different translations of loanwords

Loan translation is a method of secondary interlingual term formation (Sager, 1990: 82). When a term does not exist in Turkish some different translations are used together in terminology of Turkish lexicography so in this case polyonymy reveals. Different translation examples determined in the TLC are listed in Table 3 and they are written in italics.

Term (Text(s) / Frequency)
bilgisayar destekli sözlük bilimi (2/10), *bilgisayarlı* sözlük bilimi (2/4), *bilgisayar destekli* sözlükbilim (1/2), *bilgisayar destekli* sözlükçülük (1/1), *bilgisayarlı* sözlükbilimi (1/1), *bilgisayar destekli* sözlük yazımı (1/1)
bütüncül yapı (5/24) *dış yapı* (4/5), makro yapı (3/3), *büyük yapı* (1/2), makroyapı (1/1), *dışyapı* (1/1)
derlem tabanlı (8/18), *derlem temelli* (3/3), *derlemtemelli* (2/5), *derlemtabanlı* (1/3)
madde (113/918), *madde başı* (79/442), *sözlükbirim* (16/45), *maddebaşı* (15/120), *girdi* (13/28), entry (9/29), lemma (8/28), sözlük maddesi (7/13), sözlük birim (5/169), headword

(4/13), sözlükbirimi (4/12),*başsözcük* (3/10), sözlüksel birim (4/7), lexeme (4/7), sözlük birimi (1/11), *giri* (1/9), leksem (1/3)
metaleksikografi (3/12), metasözlükçülük (1/41), *üst sözlükçülük* (1/2)
uzmanlık alanı sözlüğü (8/16), *özel alan* sözlüğü (3/6), *uzmanlık* sözlüğü (2/8), *uzmanlık alan* sözlüğü (1/1), *özel amaca dayalı* sözlük (1/1), uzman sözlük (1/1)
standartlaşma (4/5), *ölçünleşme* (1/1)

Table 3: Polyonyms based on different translations of loanwords

2.4. Different derivational suffixes (Different derivational forms)

Derivation or affixation, which is the addition of affixes is a common method of designation of new concepts (Sager, 1990: 72). Turkish is an agglutinative language. Using derivational suffixes in Turkish is the most common word formation method. Functionally, derivation and compounding serve the purpose of closer determination of a concept-narrowing its intension-while at the same time showing the relationship that exists between the new concept and its origin. (Sager, 1990: 73). Different same or similar functioning suffixes are added to the same root in term formation in Turkish. There are many polyonym terms in Turkish lexicography that are formed in this way. Examples derivated by this way determined in the TLC are listed in Table 4 and different suffixes are written in italics.

Term (Text(s) / Frequency)
anadili konuş*ucusu* (2/4), anadili konuş*uru* (1/1), anadili konuş*anı* (3/3)
anlam değiş*mesi* (15/53), anlam değiş*ikliği* (5/6), anlam değiş*imi* (1/6), anlam*sal* değiş*im* (1/2), semantik değiş*me* (1/1), semantik değiş*iklik* (1/1)
anlam genişlemesi (10/34), anlam geliş*imi* (1/1), anlam geliş*mesi* (1/1)
ansiklopedi madde*si* (2/2), ansiklopedi*k* girdi (1/1), ansiklopedi*k* madde (1/1)
bas*ılı* sözlük (12/22), bas*kı* sözlük (1/1), bas*ılı* geleneksel sözlük (1/1), bas*ılı* kağıt sözlük (1/1), kağıt bas*ım* sözlük (1/1)
bilgisayar destekli sözlük bilimi (2/10), bilgisayar*lı* sözlük bilimi (2/4), bilgisayar destekli sözlükbilim (1/2), bilgisayar destekli sözlükçülük (1/1), bilgisayar*lı* sözlükbilimi (1/1)
çapraz gönder*me* (2/2), çapraz gönder*im* (1/5), çapraz referans (1/2)
çok*luk* şekil (3/4), çoğ*ul* şekil (2/3), çoğ*ul* biçim (2/2), çok*luk* biçim (1/1)
çok şekilli ses*lilendirme* (1/2), çok şekilli ses*lendirme* (1/1)
durum çerçeve sözlüğü (1/3), durum çerçeve*si* sözlüğü (1/1)
eş anlamlı (24/58), eşanlamlı (18/25), anlam*daş* (8/8), sinonim (5/10), müteradif (3/3)
çok*lu* yazımlı (2/21), çok yazımlı (1/1)
kat*ılımlı* sözlük (1/1), kat*kılı* sözlük (1/1)

thesaurus (5/26), kavram*lar* dizini (6/12), kavram dizini (3/3)
küçük yapı (1/3), parça*cıl* yapı (4/10), parçayapı (1/1), mikro yapı (3/3), mikroyapı (3/3)
maddebaşı*lama* (1/1), madde*leştirme* (1/1)
makine*yle* okunabilir sözlük (1/7), makine*ce* okunur sözlük (1/1)
mecaz*i* anlam (3/3), mecaz anlam (2/2)
sözlükbilim (34/167), sözlükbilim*i* (30/269), sözlük bilim*i* (20/78), sözlük bilim (9/15), leksikografi (7/17), leksikografya (2/2), sözlük*çülük* (64/458), leksikoloji (7/8), lügat*çilik* (5/7), sözlükbilgisi (5/5), sözlük bilgisi (1/1), sözlük*çülük* bilimi (1/1)
adbilim*sel* sözlük (2/2), adbilim sözlüğü (1/1), onomastik sözlük (1/1)
ödünç*leme* (5/10), ödünç kelime (2/2), ödünç sözcük (1/1), ödünç*lenmiş* sözcük (1/1)
öğren*ci* sözlüğü (3/23), öğren*ici* sözlüğü (3/4)
uzmanlık alan*ı* sözlüğü (8/16), özel alan sözlüğü (3/6), uzmanlık sözlüğü (2/8), uzmanlık alan sözlüğü (1/1), özel amaca dayalı sözlük (1/1), uzman sözlük (1/1)
ses*teş* (6/13), eş sesli (6/18), eşsesli (5/5)
sözlük*sel* (39/286), leksikal (6/11), sözlük*lük* (2/3)
sözcük*selleşme* (1/1), leksikalizasyon (1/1), sözcük*leştirme* (1/1), sözcükbirim*leştirme* (1/1)
sözlük hazırla*yıcısı* (19/38), sözlük*çü* (13/28), sözlük yazarı (13/16), sözlük derleyicisi (4/11), sözlük hazırla*yıcı* (1/1), sözlük düzenleyicisi (1/1), sözlük yapıcı (1/1)
sözlük kullan*ıcısı* (23/33), sözlük kullan*ıcı* (2/2), sözlük okuru (1/1)
sözlük*leştirme* (1/1), sözlük*selleştirme* (3/4), sözlükbirim*selleştirme* (1/1)
sözlük*selleşme* (5/95), sözlük*leşme* (1/1), leksikalizasyon (1/1)
sözlük*sel* anlam (6/10), sözlük*lük* anlam (1/2)
tarih*sel* sözlük (15/22), tarih*i* sözlük (2/2)
çeviri sözlük (3/14), tercüme sözlük (2/3), çeviri sözlü*ğü* (1/1)
ters*ine* sözlük (7/7), ters dizim sözlüğü (1/3), ters dizimli sözlük (1/1)
türetim eki (3/5), yapım eki (6/14), türetim biçimbirimi (1/1), türetim*lik* (1/1)
değiş*ken* (20/35), değiş*ke* (13/16), varyant (9/20)
yönet*ici* paneli (1/11), admin paneli (1/5), yönet*im* paneli (1/1)
zihin sözlüğ*ü* (1/9), zihin*sel* sözlük (2/26)

Table 4: Polyonyms based on different derivational suffixes (different derivational forms)

2.5. Terms derivation from different roots

Different terms designating the same concept can be derivated from different roots which have same or similar meaning (synonym roots) in Turkish. The authors may have prefered different roots based on different features of the concept in this type term formation. In this case, there are many polyonym terms derived from different roots designating the same concept in terminology of Turkish lexicography.

Term (Text(s) / Frequency)
alıntı (60/178), ödünçleme (10/18)
anadili konuşanı (3/3), anadili konuşucusu (2/4), anadili konuşuru (1/1), anadili kullanıcısı (1/1)
anlam genişlemesi (10/34), anlam gelişimi (1/1), anlam gelişmesi (1/1)
anlam ilişkisi (11/19), anlam ilgisi (3/5)
anlam daralması (3/9), anlam kısılması (1/1)
anlık oluşum (2/13), anlık üretim (1/1), anlık türertim (1/1)
ansiklopedi maddesi (2/2), ansiklopedik girdi (1/1), ansiklopedik madde (1/1)
artsüremli yöntem (1/1), artzamanlı yöntem (1/1)
başvuru kaynağı (14/20), başvuru kitabı (4/7), başvuru eseri (1/1)
çokluk şekil (3/4), çoğul şekil (2/3), çoğul biçim (2/2), çokluk biçim (1/1)
madde (113/918), madde başı (79/442), sözlükbirim (16/45), maddebaşı (15/120), girdi (13/28), entry (9/29), lemma (8/28), sözlük maddesi (7/13), sözlük birim (5/169), headword (4/13), sözlükbirimi (4/12),başsözcük (3/10), sözlüksel birim (4/7), lexeme (4/7), sözlük birimi (1/11), giri (1/9), leksem (1/3),
temel anlam (8/19), ilk anlam (5/5), birincil anlam (2/2), düzanlam (2/6), düz anlam (2/3), yaygın anlam (2/2), asıl anlam (1/5), öz anlam (1/1), ilkel anlam (1/1)
parçabütün ilişkisi (4/4), parça bütün ilişkisi (2/9), parçabütün ilgisi (1/1)
sözcük buldurucu (1/1), dizin erişim kodu (1/1), izleti (1/1), erişim kılavuzu (1/1), dizi başı (1/1), dizibaşı (1/1), dizi sonu (1/1), dizisonu (1/1), erişim hecesi (1/1), sayfa içerik başı (1/1), sayfa içerik sonu (1/1), adres (1/1), catchword (1/19)
kelime türetimi (3/3), sözcük üretimi (1/1), sözcük türetimi (1/1)
sözlük hazırlama (23/49), sözlük yazma (13/21), sözlük yazımı (9/22), sözlük yapımı (3/10), sözlük düzenleme (3/4), sözlük derleme (3/3), sözlük yapımcılığı (3/3)
sözlük hazırlayıcısı (19/38), sözlükçü (13/28), sözlük yazarı (13/16), sözlük derleyicisi (4/11), sözlük hazırlayıcı (1/1), sözlük düzenleyicisi (1/1), sözlük yapıcı (1/1)
sözlük kullanıcısı (23/33), sözlük kullanıcı (2/2), sözlük okuru (1/1)

sözlüksel boşluk (1/14), sözcüksel boşluk (1/5)
tanımlama yöntemi (6/9), tanımlama biçimi (2/3)
türetim eki (3/5), yapım eki (6/14), türetim biçimbirimi (1/1), türetimlik (1/1)
yan anlam (9/36), yananlam (3/4), ikincil anlam (1/1)
yerleşikleşme (2/24), kurumsallaşma (2/2), özelleşme (1/1)

Table 5: Polyonyms based on terms derivation from different roots

2.6. Nativization aim (Effect of the Turkish language reform)

A nativization process of borrowings relies on the alteration of phonological, morphological and orthographical forms in conformity with systems in the target language (Sager, 1990: 85).

After the foundation of the Turkish Republic in 1923 Turkish language reform has occurred in Turkey. One of the aims of the Turkish language reform is to use the Turkish origin words instead of the words that entered Turkish from languages such as Arabic and Persian. In this context, instead of the foreign terms included in the terminology of Turkish lexicography, sometimes the researchers have taken care to use their Turkish origin form. In this case, it has been observed that foreign originated terms are used in the field of Turkish lexicography and their Turkish equivalent are used at the same time.

Term (Text(s) / Frequency)
ansiklopedik *sözlük* (11/12), ansiklopedik *lügat* (1/1)
sözlükçülük geleneği (5/8), *lügatçilik* geleneği (1/1)
arkaik *kelime* (3/6), arkaik leksikon (1/1), arkaik *sözcük* (1/1)
arkaik (13/30), *eskicil* (2/2)
çekirdek sözvarlığı (2/2), çekirdek *sözcük* (1/3), çekirdek *kelime* (1/1), çekirdek söz (1/1), çekirdek söz varlığı (1/1)
çok *anlamlı* (20/48), çok*anlamlı* (8/43), çok *manalı* (1/1)
derlem (26/234), *bütünce* (7/34), *korpus* (1/1)
düzeltme *imi* (2/6), düzeltme *işareti* (1/9)
eş anlamlı (24/58), eşanlamlı (18/25), anlamdaş (8/8), sinonim (5/10), *müteradif* (3/3)
eş sesli *kelime* (3/5), eşsesli *kelime* (2/2), eşsesli sözcük (2/2), eş sesli sözcük (1/2)
zıt anlamlı (12/41), *karşıt* anlamlı (9/11), *ezdad* (2/10), antonim (1/2), zıtanlamlı (1/1)
zıt anlam (4/9), *karşıt* anlam (4/6)
zıt anlamlılık (6/11), *karşıt* anlamlılık (2/7), zıtanlamlılık (1/1)
zıtanlamlı ögeler sözlüğü (1/1), zıt anlamlı *kelimeler* sözlüğü (1/1), zıt anlamlı *sözcükler* sözlüğü (1/1), zıt anlamlı sözlük (2/2)

gerçek *mana* (2/5), gerçek *anlam* (1/2)
edat (27/115), *ilgeç* (7/84)
ilgeç öbeği (1/2), *edat* öbeği (1/1)
yazım kılavuzu (6/24), *imla* kılavuzu (4/14)
ad (265/4148), isim (135/1215)
açıklamalı sözlük (5/15), *izahlı* sözlük (2/4), *izahlı* lügat (2/2), *izahlı* lüğet (1/1)
sözlük (222/11951), *kamus* (31/100), *lugat* (60/380), *lügat* (114/688), *tuhfe* (9/92), *mucem* (7/34), tılcıt (2/2), sözdik (1/1)
kelime (209/5208), *sözcük* (176/3441)
sözcük anlamı (5/8), *kelime* anlamı (2/2)
manzum *sözlük* (5/18), manzum *lügat* (1/4)
marka *adı* (2/27), marka *ismi* (1/2)
ek (223/2082), *biçimbirim* (18/119), biçimbirimi (5/14), *morfem* (5/13), biçim birimi (2/2)
sözlükçe (11/32), *glossary* (8/10), *lügatçe* (5/6), *lugatçe* (2/14)
sözlükbilim (34/167), sözlükbilimi (30/269), sözlük bilimi (20/78), sözlük bilim (9/15), *leksikografi* (7/17), *leksikografya* (2/2), sözlükçülük (64/458), leksikoloji (7/8), *lügatçilik* (5/7), sözlükbilgisi (5/5), sözlük bilgisi (1/1), sözlükçülük bilimi (1/1)
özel *ad* (6/18), özel *isim* (6/8)
kelime türetimi (3/3), *sözcük* üretimi (1/1), *sözcük* türetimi (1/1)
kişi adı (14/17), *şahıs ismi* (3/4), *kişi ismi* (3/3), *şahıs adı* (2/2)
çeviri sözlük (3/14), *tercüme* sözlük (2/3), *çeviri* sözlüğü (1/1)
yabancı *kelimeler* sözlüğü (3/15), yabancı *sözcükler* sözlüğü (1/3)

Table 6: Polyonyms based on nativization aim

2.7. Preferences of the authors

Some researchers use original terms sometimes because of various reasons as polysemy in terms, terms does not correspond with concept, the possibility of confusing with other terms, disapprove the existing translation/s etc. Therefore, the using of a new original term to refer to the same concept with term or terms that already exist by researchers in the field can cause polyonymy.

Term (Text(s) / Frequency)
bütüncül yapı (5/24) dış yapı (4/5), makro yapı (3/3), büyük yapı (1/2), makroyapı (1/1), dışyapı (1/1)
derlem (26/234), bütünce (7/34), korpus (1/1)
derlem dilbilim (3/10), derlem dil bilimi (1/2), derlemdilbilim (1/1), bütünce dil bilimi (1/1)
yeni sözcük (11/32), yeni kelime (11/17), yeni söz (3/4), yeni sözbirim (1/1), neoloji (1/1), neology (1/1), neologie (1/1)

Table 7: Polyonyms based on preferences of the authors

2.8. Dialect

Cabré (1999: 110) points out that two designations are synonymous only in a very narrow linguistic sense and are not synonymous in a pragmatic sense. There are many cases of two synonymous units that belong to two different registers of the same language, but this does not usually appear in a single specialized dictionary. This dissymmetry occurs in cases like the following: a. Between a scientific name and its popular name; b. Between a standard form and dialectal forms.

In a special field terminology there can be some terms taken from dialects This is one of the ways of term formation. Nevertheless, in some cases, existing of dialectal word in a special field can be occured polyonymy.

Term (Text(s) / Frequency)
açıklamalı sözlük (5/15), izahlı sözlük (2/4), izahlı lügat (2/2), izahlı *lüğet* (1/1)
sözlük (222/11951), kamus (31/100), lugat (60/380), lügat (114/688), tuhfe (9/92), mucem (7/34), *tılcıt* (2/2), *sözdik* (1/1)

Table 8: Polyonyms based on dialect

2.9. Ellipsis

If an abbreviation of a phrase occurs, this is called an ellipsis (Traugott, 2005: 40). Ellipsis leads to the formal reduction of a complex word or phrase (Blank, 2001: 1605). Traugott cites that in ellipsis the semantics of the omitted element is absorbed into the remainder by metonymy; thus in an earlier work ellipsis was often regarded as a primarily semantic change (Ullmann, 1962). However, ellipsis clearly involves both form and meaning (Hock and Joseph, 1996: 175). If ellipsis is used in a special field and two or more terms continue to exist polyonymy reveals in that field. Ellipsis examples determined in the TLC are listed in Table 9 and the abbreviated forms are written in italics.

Term (Text(s) / Frequency)
tersine sözlük (7/7), ters dizim sözlüğü (1/3), ters dizimli sözlük (1/1)
türetim eki (3/5), yapım eki (6/14), türetim biçimbirimi (1/1), *türetimlik* (1/1)
yazma sözlük (2/4), el yazması sözlük (1/1)

Table 9: Polyonyms based on ellipsis

2.10. Initialisms and abbreviations

Initialisms are units made up of the combination of the initials of a longer expression. They often correspond to the name of an organization, document or process, and in many cases they become lexicalized in the common language (Cabré, 1999: 86). Initialisms are also used in terminology beside common language. We determine *esözlük* and *e-sözlük* polyonyms formed by initialism in terminology of Turkish lexicography.

Abbreviations are forms that are usually established by consensus. They reproduce a part of a word and practically act as a symbol for the word (Cabré, 1999: 87). Because of it is used an abbreviation, *IPA* term cause polyonymy in terminology of Turkish lexicography.

Term (Text(s) / Frequency)
esözlük (3/28), *e-sözlük* (3/28), elektronik sözlük (17-58)
IPA (4/8), Uluslararası Fonetik Alfabe (1/2), Uluslararası Sesbilgisi Alfabesi (1/1)

Table 10: Polyonyms based on initialisms and abbreviations

2.11. Definition type terms

Longer terms have been formed for the same concept for representing the details of the concept in term formation. We can call "definition type terms" for these type terms. If these terms are used in special fields, polyonymy reveals there.

Term (Text(s) / Frequency)
basılı sözlük (12/22), baskı sözlük (1/1), *basılı geleneksel* sözlük (1/1), *basılı kağıt* sözlük (1/1), *kağıt basım* sözlük (1/1)
bilgisayar destekli sözlük bilimi (2/10), *bilgisayarlı* sözlük bilimi (2/4), *bilgisayar destekli* sözlükbilim (1/2), *bilgisayar destekli* sözlükçülük (1/1), *bilgisayarlı* sözlükbilimi (1/1), *bilgisayar destekli* sözlük yazımı (1/1)
sözlükbilimci (17/41), sözlük bilimci (6/33), *sözlükbilim uzmanı* (1/1), leksikograf (1/1)
sözlük hazırlayıcısı (19/38), sözlükçü (13/28), *sözlük yazarı* (13/16), *sözlük derleyicisi* (4/11), *sözlük hazırlayıcı* (1/1), *sözlük düzenleyicisi* (1/1), *sözlük yapıcı* (1/1)
sözlükten çıkarma (2/2), *sözlükten madde silme* (1/1)
tersine sözlük (7/7), *ters dizim* sözlüğü (1/3), *ters dizimli* sözlük (1/1)
yazma sözlük (2/4), *el yazması* sözlük (1/1)

Table 11: Polyonyms based on definition type terms

2.12. Changing component order in a possessive construction

In some cases polyonymy reveals through changing component order in possessive construction. One example is determined in the TLC which occured by this way.

Term (Text(s) / Frequency)
mürsel mecaz (1/2), mecaz-ı mürsel (1/2)

Table 12: Polyonyms based on changing component order in a possessive construction

3. Conclusion

In this study examples of polyonymy in terminology of Turkish lexicography are examined. Turkish Lexicography Corpus was used to determine the examples. A large number of polyonym terms have been found in the corpus. There are some reasons of polyonymy. For these reasons, the topics have been classified in the study. These topics are orthographical differences, borrowings, different translations of loanwords, different derivational suffixes (different derivational forms), terms derivation from different roots, nativization aim (effect of the Turkish language reform), preferences of the authors, dialect, ellipsis, initialisms and abbreviations, definition type terms and changing component order in a possessive construction. The most common reason of polyonymy in terminology of Turkish lexicography is orthographical differences.

There are a total of 1616 terms in the Turkish Lexicography Corpus. We determine 539 polyonym terms in the corpus. The ratio of polyonym terms to the total number of terms is 33.35%. If there was no polyonymy, there would be 1250 terms in the corpus. 173 of the 539 terms remain, unless we consider polyonym terms. 366 terms were not in the corpus if one concept corresponded to one term. These terms can be seen as a redundancy. The ratio of redundancy due to polyonymy to the total number of terms is 22.64%.

References

Blank, A. (2001). "Pathways of lexicalization". In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, eds., *Language Typology and Language Universals*, Vol. II, 1596–1608. (Handbücher zur Sprach- und Kommunikationswissenschaft, 20.2.) Berlin and New York: Walter de Gruyter.

Brinton, L. J. and Traugott, E. C. (2005). *Lexicalization and Language Change*, New York: Cambridge University Press.

Boz, E., Bozkurt, F. and Doğru F. (2017). "Türk Sözlükbilimi Terminolojisi Üzerine Derlem Tabanlı Bir Araştırma", *III. Uluslararası Sözlükbilimi Sempozyumu Bildiri Kitabı*. Eskişehir: Eskişehir Osmangazi University Press.

Cabré, M. T. (1999). *Terminology, Theory, Methods and Applications*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hock, H. H. and Joseph, B. D. (1996). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics.* Berlin and New York: Mouton de Gruyter.

Murphy, M. L. (2003). *Semantic Relations and the Lexicon*. New York: Cambridge University Press.

OED Oxford English Dictionary 2018. London: Oxford University Press, viewed 09 April 2018, <https://en.oxforddictionaries.com/definition/polyonym>.

Sager, J.C. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Temmerman, R. (2000). *Towards New Ways of Terminology Description, The Sociocognitive-Approach*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ullmann, S. (1962). *Semantics: An Introduction to the Science of Meaning*. New York: Barnes and Noble.

Vogel, R. (2006). *Lexical Hierarchies in the Scientific Terminology*, Masaryk University Faculty of Arts, Department of English and American Studies Doctoral Thesis, Brno.

MONITORING ACADEMIC STUDIES OF TURKISH LEXICOGRAPHY: A PHOTOGRAPH OF 84 YEARS

Ferdi BOZKURT, PhD
Anadolu University
ferdib@anadolu.edu.tr

Abstract

Mahmut Kashgar compiled the first dictionary of Turkish language. He completed Divânu Lügati't-Türk (the dictionary) on February 1074. Turkish lexicography has got a long tradition spanning over centuries; however, it is found to be deficient in many aspects including the realm of theoretical studies which were still not adequate.

The establishment of the Turkish Language Institute, which was established in 1932, triggered linguistic studies such as morphology, syntax, lexicology, phonetics, semantic, and etymology. As a result of these studies, books, doctoral dissertations, articles, papers, presentations and many other studies have been published and continue to be published. There is no academic journal which relates only to Turkish lexicography in Turkey, however, there are many academic journals including grammar and linguistics research studies. Topics related to Turkish lexicography are generally published in the journals of institutes and/or linguistics and grammar journals. Since the number of studies on Turkish language is too many and varied, it was necessary to distinguish the studies related to Turkish lexicography among the studies conducted in different linguistics subfields.

The aim of this study is to reveal the number of the texts written in the field of Turkish lexicography according to the variables listed above and to present trends in the field of Turkish lexicography in 84 years.

Keywords: Turkish lexicography, articles, presentations, master theses, doctoral dissertations

1. Introduction

The first dictionary study known for Turkish language began with Mahmut Kashgar. He completed Divânu Lügati't-Türk (the dictionary) on February 1074 (Ercilasun, 2015). Turkish lexicography has got a long tradition spanning over centuries; however, it is found to be deficient in many aspects including the realm of theoretical studies which were still not adequate (Boz, 2006).

The Turkish Language Institute, which was established in 1932, triggered linguistic studies such as morphology, syntax, lexicology, phonetics, semantic, and etymology. As a result of these studies, books, doctoral dissertations, articles, presentations, bulletins and many other studies have been published and continue to be published. There is no academic journal which relates only to Turkish lexicography in Turkey, however, there are many academic journals including grammar and linguistics research studies. The topics related to Turkish lexicography are generally published in the journals of institutes and/or linguistics and grammar journals. Since the number of studies on Turkish language is too many and varied, it was necessary to distinguish the studies related to Turkish lexicography among the studies conducted in different linguistics subfields.

The number of bibliographic studies in the field of Turkish lexicography is rather limited. In this field, generally, bibliographical studies about the prepared dictionaries and the reviews of these dictionaries are being made (Eminoglu 2010; Kotan, 2017).

The study of Yıkmış and Sazak (2017) determines 9 books, 31 book chapters, 106 thesis, 188 presentations, 23 reviews and 219 articles related to lexicography. The studies of Yıkmış and Sazak (2017) have been limited between the years 2000 and 2016. These two researchers have divided the texts into text types like books, book chapters, doctoral dissertations, master theses, presentations, reviews and articles.

All the texts related to the field of Turkish lexicography between 1932 and 2016 were gathered and a database was prepared for the project of the Turkish Lexicography Corpus, in which I was a researcher (Boz, et al, 2017).

2. Method

In this section, information about the research model, analyzed documents, the data collection tool and the analysis of the data are mentioned. In order to create a database to be used in the current study, the field of Turkish lexicography literature was reviewed. Both printed and non-printed books, academic journals and electronic databases of thesis were reviewed for the current study.

In Turkey, there are thousands of academic studies in the field of morphology, syntax, phonetics, phonology, psycholinguistics, sociolinguistics, computational linguistics, historical linguistics, applied Linguistics etc. There are thousands of books, articles, doctoral dissertations and master theses on linguistics studies. In order to create the database to be used in the research, while the scanning is being done by physical libraries, ULAKBİM, Google Scholar, Turkey National Thesis Center, TO-KAT Turkish National Library, EbscoHost, Proquest Dissertations and Theses Global databases were used.

As the boundaries of field of Turkish lexicography are not clear cut, a criterion was defined to select the text for the database. The criterion was to filter the texts whether they had some specific keywords; "sözlük" (dictionary), "lügat" (dictionary, an old usage), "sözlükbilim" (lexicography), "sözlük bilim" (lexicography), "sözlükbilimi" (lexicography), "sözlük bilimi" (lexicography), "sözlükçülük" (synonym with lexicography), "leksikografi" (lexicography) were included in the database (Boz et. al. 2017). A total of 1001 texts which were written between 1932-2016 were identified as a result of this search.

Table 1: Text types determined for the current study

Text Type	Number(s) of Texts	%
Master theses	39	3.90
Doctoral dissertations	12	1.20
Published presentations	301	30.07
News	21	2.10
Books	3	0.30
Articles	475	47.45
Reviews	150	14.99
Total	1001	100.00

3. Findings

1001 texts were analyzed to determine the research tendency in this field by analyzing the published texts in the field of Turkish lexicography between 1932 and 2016. Frequency and percentage values for text year of publication are shown in Table 2:

Table 2: Number(s) of publication by years

Year	Number(s)	%	Year	Number(s)	%	Year	Number(s)	%
1934	1	0.10	1973	2	0.20	1996	9	0.90
1935	2	0.20	1974	2	0.20	1997	5	0.50
1936	3	0.30	1975	13	1.30	1998	15	1.50
1939	2	0.20	1976	5	0.50	1999	31	3.10
1942	1	0.10	1977	6	0.60	2000	16	1.60
1952	2	0.20	1979	1	0.10	2001	5	0.50
1953	3	0.30	1980	2	0.20	2002	18	1.80
1954	6	0.60	1981	2	0.20	2003	8	0.80
1956	2	0.20	1982	3	0.30	2004	22	2.20
1957	1	0.10	1983	1	0.10	2005	12	1.20
1959	3	0.30	1984	1	0.10	2006	21	2.10
1960	3	0.30	1985	2	0.20	2007	63	6.29
1961	1	0.10	1986	2	0.20	2008	84	8.39
1962	2	0.20	1987	2	0.20	2009	111	11.09
1963	1	0.10	1989	1	0.10	2010	59	5.89
1965	2	0.20	1990	1	0.10	2011	68	6.79
1967	2	0.20	1991	4	0.40	2012	46	4.60
1968	1	0.10	1992	1	0.10	2013	70	6.99
1969	2	0.20	1993	1	0.10	2014	52	5.19
1970	1	0.10	1994	5	0.50	2015	61	6.09
1971	4	0.40	1995	8	0.80	2016	98	9.79
1972	17	1.70	**Total**				1001	100.00

As it can be seen in the table, the most producing texts in Turkey are about lexicography; 2009 (111 texts), 2016 (98 texts), 2008 (84 texts), 2013 (70 texts), 2011 (68 texts.). In some years (1937, 1938, 1940, 1941) texts related to the field of Turkish lexicography did not produce any text. However, the texts have been produced annually and uninterruptedly on the field of lexicography since 1967. The text production average is 15.6 in

the years that the text is produced. This study covers the years between 1932 and 2016, the average number of texts is 11.9 in these years.

As shown in Table 2, only one text was produced in some years (1983, 1984, 1968 etc.). And also, as it can be seen in the table, an increase has been observed in the number of texts related to lexicography over the last 20 years.

Table 3: Master theses and doctoral dissertations by years

Master theses by years

Year	Number(s)	%
1996	1	2.56
1997	1	2.56
2005	1	2.56
2006	2	5.13
2007	6	15.38
2008	1	2.56
2009	3	7.69
2010	4	10.26
2011	4	10.26
2012	5	12.82
2013	4	10.26
2014	2	5.13
2015	5	12.82
Total	39	100.00

Doctoral dissertations by years

Year	Number(s)	%
1993	1	8.33
2000	1	8.33
2007	1	8.33
2009	2	16.67
2010	3	25.00
2011	2	16.67
2013	1	8.33
2016	1	8.33
Total	12	100.00

Master theses seem to have increased in recent years shown by table 3. The numbers of doctoral dissertations are generally insufficient. The numbers of master theses are more than the numbers of doctoral dissertations in the field of Turkish lexicography.

Table 4: Presentations by years

Year	Number(s)	%	Year	Number(s)	%
1969	1	0.33	2007	37	12.29
1972	2	0.66	2008	60	19.93
1985	2	0.66	2009	36	11.96
1999	9	2.99	2010	23	7.64
2000	4	1.33	2011	7	2.33
2002	1	0.33	2012	9	2.99
2003	1	0.33	2013	25	8.31
2004	8	2.66	2015	1	0.33
2005	1	0.33	2016	72	23.92
2006	2	0.66	**Total**	301	100.00

The total number of presentations is one of the most produced text types compared with other text types. It is observed that the numbers of presentations have increased in the last decade, but in some years the numbers of the presentations have been quite low compared to other years. One of the most important essential factors for the increasing numbers of

presentations is lexicographical symposiums which have been organized in Turkey in recent years.

Table 5: News by years

Year	Number(s)	%	Year	Number(s)	%
1952	1	4.76	1999	2	9.52
1953	1	4.76	2004	1	4.76
1959	2	9.52	2007	2	9.52
1960	1	4.76	2009	1	4.76
1962	1	4.76	2011	2	9.52
1971	1	4.76	2012	2	9.52
1987	1	4.76	2013	1	4.76
1989	1	4.76	2014	1	4.76
Total				21	100.00

The number of news related to lexicography is produced irregularly by years.

Table 6: Books by years

Year	Number(s)	%
2007	1	33.3
2011	1	33.3
2016	1	33.3
Total	3	100.00

In the field of Turkish lexicography, the number of books is insufficient. There is still no handbook of lexicography on Turkish lexicography as the Europeans' examples (Durkin, 2016; Svensén, 2009; Atkins and Rundell, 2009; Jackson, 2002; Zgusta, 1971).

Table 7: Articles by years

Year	Number(s)	%	Year	Number(s)	%	Year	Number(s)	%
1934	1	0.21	1974	1	0.21	2001	4	0.84
1939	1	0.21	1977	1	0.21	2002	11	2.32
1942	1	0.21	1981	1	0.21	2003	6	1.26
1952	1	0.21	1983	1	0.21	2004	11	2.32
1953	2	0.42	1984	1	0.21	2005	5	1.05
1954	1	0.21	1986	2	0.42	2006	14	2.95
1956	1	0.21	1987	1	0.21	2007	14	2.95
1957	1	0.21	1990	1	0.21	2008	18	3.79
1959	1	0.21	1991	2	0.42	2009	64	13.47
1960	1	0.21	1994	3	0.63	2010	22	4.63
1961	1	0.21	1995	3	0.63	2011	46	9.68
1965	1	0.21	1996	3	0.63	2012	24	5.05
1970	1	0.21	1997	4	0.84	2013	36	7.58
1971	2	0.42	1998	15	3.16	2014	39	8.21
1972	9	1.89	1999	16	3.37	2015	48	10.11
1973	1	0.21	2000	9	1.89	2016	23	4.84
Total							475	100

The most producing text type in the field of Turkish lexicography is article. The ratio of the total text type of the article among the types of texts produced in this area is 47.4%. 475 articles were produced in 48 total years. In the field of Turkish lexicography, an average of 9.8 articles is produced annually. The numbers of articles have shown an upward trend in the last 10 years.

Table 8: Reviews by years

Year	Number(s)	%	Year	Number(s)	%	Year	Number(s)	%
1935	2	1.33	1975	13	8.67	2002	6	4.00
1936	3	2.00	1976	5	3.33	2003	1	0.67
1939	1	0.67	1977	5	3.33	2004	2	1.33
1954	5	3.33	1979	1	0.67	2005	5	3.33
1956	1	0.67	1980	2	1.33	2006	3	2.00
1960	1	0.67	1981	1	0.67	2007	2	1.33
1962	1	0.67	1982	3	2.00	2008	5	3.33
1963	1	0.67	1991	2	1.33	2009	5	3.33
1965	1	0.67	1992	1	0.67	2010	7	4.67
1967	2	1.33	1994	2	1.33	2011	6	4.00
1968	1	0.67	1995	5	3.33	2012	6	4.00
1969	1	0.67	1996	5	3.33	2013	3	2.00
1971	1	0.67	1999	4	2.67	2014	10	6.67
1972	6	4.00	2000	2	1.33	2015	7	4.67
1973	1	0.67	2001	1	0.67	2016	1	0.67
1974	1	0.67	**Total**				150	100

According to years, the numbers of reviews are close to each other. Although the numbers of reviews have increased for some years, the numbers of reviews continue with an average number of 3.2%.

Table 9: Most producing researchers

	Name of Researcher	Number(s) of texts		Name of Researcher	Number(s) of texts
1	Erdoğan BOZ	24	11	Paşa YAVUZARSLAN	5
2	Turkish Language Institute	20	12	Nuh DOĞAN	5
3	Tuncer GÜLENSOY	12	13	Sami N. ÖZERDİM	5
4	Adem AYDEMİR	9	14	Zuhal KARGI ÖLMEZ	5
5	Mehmet ÖLMEZ	8	15	Bülent ÖZKAN	5
6	İsmail PARLATIR	8	16	Akartürk KARAHAN	5
7	Ali PÜSKÜLLÜOĞLU	8	17	Aysu ATA	5
8	Galip GÜNER	7	18	Atabey KILIÇ	5
9	Nail TAN	6	19	Hasan EREN	5
10	Fatih DOĞRU	6	20	Zeynep KORKMAZ	4

Twenty researchers who are the most productive on Turkish lexicography were shown in Table 9. In the database of the current study there are 678 researchers who produce the texts related to Turkish lexicography. The ratio of the total number of texts to the number of researchers is 1.4%. The first three people who are the most productive on Turkish lexicography are Erdoğan Boz, Turkish Language Institute and Tuncer Gülensoy. (Studies related to Turkish Language Institute lexicography are sometimes published by the corporate

name.) The first 20 researchers producing the most text produced 15.7% of the total text amount.

Conclusion and Discussion

In this study, the ratios of the texts produced in 84-year-period in the field of Turkish lexicography are discussed. 1001 texts were produced in 84-year-period. The texts are limited by the year between 1932 and 2016, and this caused some work on printing stage since in these years mostly the printing was not that popular. The books of International Lexicography Symposium conducted in Turkey in 2014 and 2015 are not yet published because of some problems encountered in the printing stage. This caused approximately 150 presentations to fail to be in the database of the current study.

One of the main problems in the field of lexicography in Turkey is that the insufficient number of the books. The number of books needs to be increased. There is not a handbook about Turkish lexicography yet.

Insufficient number of doctoral dissertations is also another problem. A large part of the study in the field of lexicography in Turkey is composed of studies like presentations and articles which have short time process and small volume.

There is no an academic journal which includes studies only about the Turkish lexicography in Turkey. The articles are published in various linguistic journals. This situation makes it difficult for the researchers to follow the field studies.

The only database in which the texts of lexicography were compiled from 1932 to present is the Turkish Lexicography Corpus in which I work as a researcher. TLC makes suggestions to the researchers with the help of its bibliography.

In the field of lexicography, the number of researchers who produce text is high in number, but the average numbers of texts are quite insufficient.

Future Work

The texts in the database used in the current study will be thematically marked and the thematic features of the texts in the Turkish lexicography will be revealed.

References

Atkins, B. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Boz, E. (2006). "Sözlük ve Sözlükçülük Sorunu", *Türkçenin Çağdaş Sorunları*, Ankara: Gazi Kitabevi.

Boz, E., Bozkurt, F., & Doğru F. (2017). "Türk sözlükbilimi terminolojisi üzerine derlem tabanlı bir araştırma", *III. Uluslararası Sözlükbilimi Sempozyumu Bildiri Kitabı*. Eskişehir: Eskişehir Osmangazi University Press.

Durkin, P. (2016). *The Oxford handbook of lexicography*. Oxford University Press.

Eminoğlu, E. (2010). *Türk dilinin sözlükleri ve sözlükçülük kaynakçası*. Asitan Yayıncılık.

Ercilasun, A. B. (2015). *Dîvânü Lugâti't-Türk: giriş, metin, çeviri, notlar, dizin*. Ankara: Türk Dil Kurumu Yayınları.

Jackson, H. (2002). *Lexicography: An Introduction*. Routledge.

Kotan, H. (2017). "Erzurum yazma eserler kütüphanesinde bulunan el yazması sözlükler". *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 21/4: 1491- 1509.

Svensén, B. (2009). A *handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Tietze, A. (1976). "Problems of Turkish Lexicography". *Problems in Lexicography* (Edit. by Fred W. Householder and Sol Saporta). Bloomington: Indiana University: 263-272.

Yıkmış, S., & Sazak, Ş. (2017) "Türk sözlükbilimi bibliyografyası üzerine bir deneme (2000-2016)". *III. Uluslararası Sözlükbilimi Sempozyumu Bildiri Kitabı*. Eskişehir: Eskişehir Osmangazi University Press.

Zgusta, L. (1971). *Manual of lexicography*. Walter de Gruyter.

A Filipino-English Disaster Sentiment Polarity Lexicon

Angelica Dela Cruz, Nathaniel Oco, Rachel Edita Roxas
National University – Manila, Philippines
{ahdelacruz, naoco, reoroxas}@national-u.edu.ph

Abstract

In this paper, we present a novel approach, based on an ongoing study, to cluster 43 Philippine languages towards building an updated Philippine language family tree using orthographic features. The 43 languages are classified as follows: 17 are identified as developing, 13 as educational, 10 as wider communication, 2 as threatened and 1 as vigorous. Included in the 43 languages is Yami language, spoken in Taiwan but considered similar with Ivatan, a northern Philippine language. We used orthographic features our main source of data. In particular, we used character trigrams (3-gram), which are 3-character slices of a word. For example, the word "lexicon" will produce a trigram model of {"_le", "lex", "exi", "xic", "ico", "con", "on_"}. Our corpus consists of religious text from the holy Bible. We used existing applications to generate the trigrams per language and specifically used hierarchical clustering to build the hierarchy of languages based on feature similarity among languages. Our results are comparable to the language subgroups of Ethnologue. To further our analyses, a cognate list was also used to determine language similarity. As example, words like "hapon" (afternoon), "braso" (arm) and "dugo" (blood) were found in at least two languages from the same subgroup. The work can be extended by considering other techniques and features.

Keywords: Hierarchical clustering, trigrams, Philippine languages, cognate list

1. Introduction

According to Ethnologue[1], there are 187 listed languages in the Philippines, 41 are institutional, 72 are developing, 45 are vigorous, 14 are in trouble, 11 are dying and 4 are already extinct. The existing language family tree which can be found in Ethnologue were based from different studies in order to show information about Philippine languages. The study of Wurm [11] in 2007 was used as basis to determine if the specific language is endangered or not. The study of Crystal [1] in 2003 was used as basis of including English in the list of Philippine languages because it is used globally while the studies of Reid [10], Zorc [12], and Lobel [4], [5], [6] were used as basis for presenting the relationship between Philippine languages. These studies made serious efforts in constructing subgroup of Philippine languages with the use of different features such as morphology, phonologies, syntactic features and word list to represent the languages. But as time passes by, languages have a tendency to evolve and be influenced by its neighbor languages, especially its phonetic features, and might cause the need for Philippine language subgrouping to be updated. In addition, these studies were all conducted using manual means which are laborious and time consuming. Also, these studies are focused on specific language subgroups that are only part of the Philippine language family tree.

There are existing studies that conducted experiments to automatically measure similarity between languages. These includes [7], [8] and [9] but only limited languages were covered.

[1] http://www.ethnologue.com/country/PH/languages

Previous studies [2], [3] used various features such as phonetic, orthographic and geographical features to automatically cluster languages, however, the evaluation scores were still a bit low. In order to improve its results, the goal of this initial study as part of an ongoing study is to automatically cluster 43 Philippine languages towards building a Philippine language family tree using orthographic features. Character trigram (3-gram), which are 3-character slices of a word were used in this study to represent the orthography of the languages. For example, the word "lexicon" will produce a trigram model of {"_le", "lex", "exi", "xic", "ico", "con", "on_"}. A cognate list was also used to further the analyses.

2. Methodology

A. Data collection

There are two orthographic features used in the study: trigram models and language word list. For the trigram models, online religious text documents of all the domain languages were collected. Considering the available resources gathered, the number of words used were only limited to 100,000, this is important to have fair results with all the languages gathered. The word list of the languages were given by the "Komisyon sa Wikang Filipino" that consists of 200-300 words for all the languages.

B. Data processing

The collected data were cleaned to remove all characters that are not necessary in generating trigrams such as numbers and punctuation marks. Regular expressions were utilized using Notepad++ in order to clean the data.

C. Clustering

The processed data were fed to a data mining tool to automatically cluster the languages using orthographic features. Hierarchical clustering algorithm was used in clustering the languages, it is a clustering method that builds hierarchy of groups of languages based on the similarity between the languages.

D. Evaluation

Resulting clusters will be evaluated using both extrinsic and intrinsic evaluation metric. As this is part of an ongoing study, the resulting clusters made from the orthographic features only are evaluated using an extrinsic evaluation metric first. For the extrinsic evaluation metric, we will use purity. This evaluation metric measures the validity of the clusters made by measuring the similarity of languages within a cluster based on an external expert knowledge. The ethnologue Philippine language subgrouping was used as the gold standard. Initial results are not yet evaluated using intrinsic evaluation.

3. Results and Discussion

Based on the initial results of the experiment, it can be observed that Yami and Ivatan are also found similar while Chavacano and Sama are considered as outlier languages. Our results, after evaluation, are comparable to the language subgroups of Ethnologue. To further our analyses, cognate list was also used to determine language similarity. As example, words like "hapon" (afternoon), "braso" (arm) and "dugo" (blood) in the Tagalog language are also present in the Cebuano language. Both Tagalog and Cebuano are under one language subgroup. The work can be extended by considering other techniques and features. Although the results of this study is already comparable to the language subgroup of ethnologue, there is still a need to collect data on other Philippine languages not yet covered by this study. Also,

the results are still evaluated using external evaluation metric which uses external basis which may be outdated.

4. Acknowledgement

This work is supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (No. IIID-2015-07).

5. References

[1] Crystal, D. (2003). English as a global language.
[2] Dela Cruz, Angelica, Maria Cristina Co, Adrian Martin Sy, Nathaniel Oco. (2017). Building a Language Family Tree using Various Features. In Proceedings of the 17th Philippine Computing Science Congress.
[3] Dela Cruz, Angelica, Nathaniel Oco, Leif Romeritch Syliongka, Rachel Edita Roxas. (2016). Phoneme Inventory, Trigrams, and Geographic Location as Features for Clustering Different
[4] Lobel, J. W. (2004). Old Bikol-um-vs. mag-and the loss of a morphological paradigm. *Oceanic Linguistics*, *43*(2), 469-497.
[5] Lobel, J. W. (2005). The angry register of the Bikol languages of the Philippines. *Liao and Rubino*, 149-166.
[6] Lobel, J. (2013). Philippine and North Bornean Languages: Issues in Description, Subgrouping and Reconstruction.
[7] Oco, N., Syliongka, L.R., Roxas, R.E. (2016). Clustering Philippine Languages. In: 16th Philippine Computing Science Congress (PCSC).
[8] Oco, N., Sison-Buban, R., Syliongka, L.R., Roxas, R.E., Ilao, J. (2014). Trigram Ranking: Metric for Language Similarity and Clustering. Malay, pp. 53-68
[9] Oco, N., Ilao, J., Roxas, R.E., Syliongka, L.R. (2013). Measuring Language Similarity using Trigrams. 2013. International Conference on Recent Trends in Information Technology (ICRTIT).
[10] Reid, L. (1971). Philippine minor languages: word lists and phonologies. (Oceanic Linguistics Special Publication No. 8.) xiii, 241 pp. [Honolulu]: University of Hawaii Press.
[11] Wurm, S. A. (2007). Australasia and the Pacific. encyclopedia of the world's endangered languages, 425.
[12] Zorc, D. (1977). The Bisayan Dialects of the Philippines: Subgrouping and Reconstruction. Pacific Linguistics. Series C - No. 44. The Australian National University.

A Filipino-English Disaster Sentiment Polarity Lexicon

Joseph Marvin Imperial[1], Jeyrome Orosco[2], Shiela Mae Mazo[2], Lany Maceda[2], Nathaniel Oco[1], Rachel Edita Roxas[1]
[1]National University, Philippines
[2]Bicol University, Philippines
imperialjoseph@rocketmail.com, {orosco.jeyrome, shielamae.mazo}@bicol-u.edu.ph, lanylm@yahoo.com, {naoco, reoroxas}@national-u.edu.ph

Abstract

In this paper, disaster related tweets published during the timeframe of Typhoon Yolanda (from November 1, 2013 to January 31, 2014) were data mined for the analysis of sentiments and construction of a combined Filipino and English disaster polarity lexicon. A total of 92,040 tweets were obtained using the hashtags '#YolandaPH', '#ReliefPH', '#BangonPilipinas' which were active and frequently used by the public media during Typhoon Yolanda. Further, the collected tweets were narrowed down to 39,867 after removing duplicates, retweets, and symbols. After cleaning, a small percentage of the tweets were manually annotated by an expert as ground truth and classified into three classes: positive, negative, and neutral. We trained a system to automatically perform classification and our findings reveal that 51.1% of the tweets are positive and express support, love, and words of courage to the victims; 19.8% are negative and state sadness and despair for the loss of lives and hate for corrupt officials; while the other 29% are neutral tweets from local news stations, announcements of relief operations, donation drives, and observations by citizens. The system has a reliability score of 79%. The top 15 frequently occurring words for each polarity and their frequency were identified. Included in the list of words were 'affected', 'good', 'safe' for positive class, 'missing', 'affected', 'strongest' for neutral class, and 'heartbreaking', 'affected', for the negative class. We noted that there are inter-class homographs such as 'affected' which caused misclassification of sentiments.

Keywords: *sentiment, disaster, polarity, homograph, tweets*

Introduction

The Philippines is a common birthplace of natural disasters and calamities. Due to its unfortunate geographical location along the western rim of Pacific Ocean, the Philippine Area of Responsibility (PAR) is vulnerable to frequent storm surges and formations of low pressure areas (LPA). [1] In a single year, around 20 tropical cyclones visit the country. Among these cyclones, 50% may have a chance of becoming a typhoon and 25% of these typhoons may grow as super typhoons. [2] Such natural phenomena can leave increasing death tolls, downfall in economic growth, and massive destruction of infrastructures in its wake. As typhoons grow and increase in power, so does the potential damage it can inflict upon its predicted pathway [3]. This devastation affected around 13% (12.9 million) Filipinos, 1.9 million people left homeless, 281,091 houses destroyed, and around 6021 dead. [4] Tacloban City in Leyte, Eastern Visayas was the most affected city with the heaviest damage from the typhoon which was similarly concluded by the US-based Joint Typhoon Warning Center (JTWC) as the fourth strongest super typhoon in world history.

Various sentiments, opinions, emotions and thoughts of Filipino users to different events like natural disasters can be expressed through the use of social media as this information sharing increases day by day. As such, Twitter is a social media networking service platform composed of 328 million active users and publishing almost 500 million tweets from users every day. Keen et al. developed a general Filipino-English sentiment lexicon using word level and corpus-based approach with 22,380 words [5]. For this study, a bilingual Filipino-English disaster-related sentiment polarity lexicon was constructed from the frequent words present in the sentiments of people from the Twitter social media platform during the timeframe of Typhoon Yolanda.

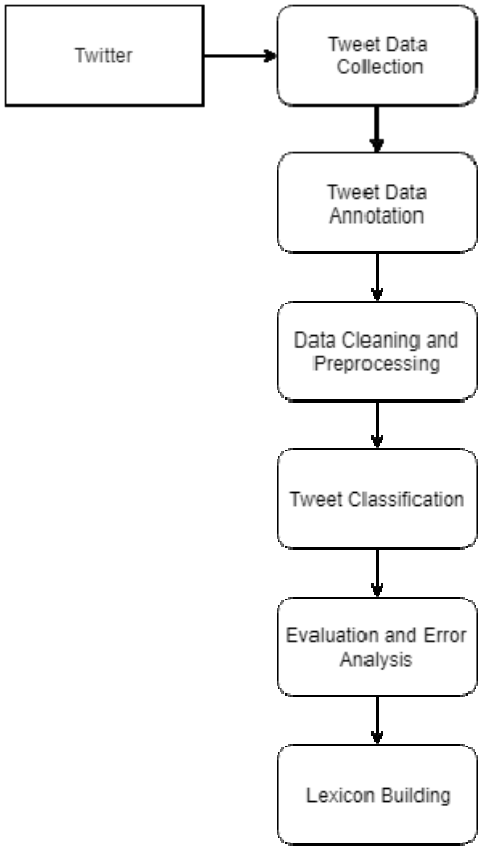


Figure 1. Disaster polarity sentiment building methodology.

Methodology

This study follows the process of analysis shown in Figure 1. Typhoon Yolanda related tweets for training and testing were gathered using the keywords '*#YolandaPH*', '*#BangonPH*', and '*#BangonPilipinas*' starting from November 1, 2013 to January 1, 2014. A total of 92,040 tweets were successfully obtained from the given time period. Duplicate tweets, usernames, external links, URLs, retweets, emojis, stop words, words appearing less than five times, words less than three characters, and special characters were removed from the existing dataset.

After preprocessing, a total of 39,867 tweets remained. From the total retrieved tweets, a random sample of 3900 tweets were selected for manual labelling. The annotated data served as the gold standard and ground truth of sentiments for the classifier. Tweets were categorized in one of the three classes: positive, negative, or neutral. Positive tweets include prayers, sympathy for the victims, and encouragement; negative tweets are those connoting stress, pity, and anger; while the neutral class contains news reports, announcements, observations by the people, plans, and tweets with no distinguishable sentiments. Supervised training was done using standard Recurrent Neural Networks. The result of classification was evaluated using accuracy as evaluation metric and analysis of errors and misclassification was noted. For building the disaster sentiment lexicon, the top 15 occurring common words as well as the top 15 adjectival words and their number of occurrences within the whole dataset were collated per polarity for analysis of similarity and frequency.

Results and Discussion

The model generated by the standard Recurrent Neural Network for fine-grained classification with an accuracy of 81.79% was used to identify the sentiments of the remaining 35,967 tweets from the original gathered Yolanda dataset. An expert validated the classified tweets for reliability of classification using three randomized test cases. The best model achieved a reliability score of 79%. 18,395 (51.1%) of the tweets were classified as positive, 10,441 (29%) tweets were neutral, and the remaining 7,131 (19.8%) tweets were negative. In the error analysis of fine-grained classification, 12% were misclassified by the final standard Recurrent Neural Network classifier model. Table 1 shows the top 15 polarity words found per class after classifying the rest of the unlabelled 35,967 tweets. Included in the list of words were 'affected', 'good', 'safe' for positive class, 'missing', 'affected', 'strongest' for neutral class, and 'heartbreaking', 'affected', for the negative class. The proponents note the frequent occurrence of inter-class homographs such as 'affected', 'missing', and 'strong' which may contribute to the misclassification of sentiments. Similarly, the proponents also tabulated the top 15 common disaster related words and their number of occurrence per polarity as shown in Table 2.

Table 1. Disaster related polarity lexicon of adjectival words and their occurrence

Positive	No. of occurence	Negative	No. of occurence	Neutral	No. of occurence
affected	484	heartbreaking	203	missing	219
good	389	affected	186	affected	215
safe	313	missing	182	free	213
strong	262	grabe	177	strongest	158

strongest	143	dead	94	good	71
missing	125	super	86	kind	57
little	121	strongest	53	local	52
many	119	safe	51	canned	52
whole	115	good	46	dead	49
loved	98	nakakaiyak ('tearful')	46	loved	43
best	96	kawawa ('poor')	44	safe	41
stronger	92	injured	37	first	39
wide	89	breaking	34	strong	39
last	86	near	32	bad	33
great	85	trending	31	updated	30

Table 2. Disaster related polarity lexicon of common words and their occurrence

Positive	**No. of occurence**	**Negative**	**No. of occurence**	**Neutral**	**No. of occurence**
help	3302	tacloban	837	help	1842
typhoon	1641	help	836	please	1578
victims	1522	news	604	tacloban	1533
philippines	1400	leyte	595	typhoon	1083
thank	983	please	531	relief	1003
survivors	757	family	476	victims	954
pray	673	typhoon	413	leyte	923
people	651	yolanda	373	city	821
please	605	relief	369	donations	759
yolanda	577	still	323	looking	705
need	535	about	322	yolanda	679
haiyan	535	city	256	family	621
love	488	bagyo ('typhoon')	224	donate	579

hope	484	know	203	need	570
after	395	wala ('none')	202	goods	504

Acknowledgement
This work is supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced Research Institutes Project (No. IIID-2015-07).

References
[1] Brown, S. (2013, November 11). The Philippines Is the Most Storm-Exposed Country on Earth. Retrieved November 21, 2017, from http://world.time.com/2013/11/11/the-philippines-is-the-most-storm-exposed-country-on-earth/
[2] Cruz, G. D. IN NUMBERS: Typhoons in the Philippines and the 2016 polls. Retrieved November 21, 2017, from https://www.rappler.com/move-ph/issues/disasters/126001-typhoons-enter-philippines-fast-facts
[3] Philippines Typhoon Facts and Figures. (2015, August 05). Retrieved November 21, 2017, from https://www.dec.org.uk/articles/facts-and-figures
[4] Quick facts: What you need to know about Super Typhoon Haiyan. (2013, November 16). Retrieved November 21, 2017, from https://reliefweb.int/report/philippines/quick-facts-what-you-need-know-about-super-typhoon-haiyan
[5] Keen, D., King, N., Lopez, J., Mondares, A., & Ponay, C. (2015). FilCon: Filipino Sentiment Lexicon Generation Using Word Level-Annotated Dictionary-Based and Corpus-Based Cross Lingual Approach. Proceedings for The Asian Association for Lexicography, 316-329. Retrieved May 24, 2018.

Pronunciation in EFL Dictionaries: A case of *during* in American English

Kensei Sugayama
Formerly Kobe City University of Foreign Studies
kensei.sugayama@uclmail.net

Abstract

The indication of pronunciation in bilingual dictionaries with English as the source language has always been an important factor especially when the other language is not at all linguistically related to the English language, primarily because the latter has a quite different phonetic and phonological system from English, let alone its different syntax and semantics. This is particularly significant for bilingual dictionaries for the learners of English.

As non-native learners of English we expect EFL dictionaries to describe the standard language --- that form of the language that is most valuable and efficacious to us. It will be understood by most of the native speakers and it is the form of the language that most facilitates international communication between non-native speakers.

Japan is said to be one of the countries which have produced the most user-friendly English-Japanese dictionaries including *Genius*, *Wisdom*, etc. Yet the recent EFL dictionaries published in Japan see a peculiar thing happening in the phonetic transcription of the word *during* for American English. Almost all the Japanese EFL dictionaries list the not-so-common /ˈdə□rɪŋ/ rather than a more usual /ˈd(j)ʊrɪŋ/ as the recommended pronunciation for American English, contrary to the fact that the native-speakers use rarely this recommended pronunciation and that most American dictionaries do not give this pronunciation under the entry of *during*. One may no doubt wonder why this has happened.

In this short article I will explain what caused this to happen in the scene of the EFL lexicography in Japan and try to suggest ways of rectifying the current situation following Professor John C. Wells's comment and advice (p.c.).

In order to answer the question of what standard pronunciation is, ask the informants whether a given pronunciation reflects how the word is pronounced or collect as much information as is required from large-scale spoken corpora.

Keywords pronunciation, American English, vowel change, descriptive linguistics

Introduction and the Problem

Quite recently I have noticed a somewhat inconsistent description in EFL dictionaries published in Japan, where English is taught as a foreign language in the secondary and tertiary education. It is surprising to know this fact, which this paper explores in detail below, has never been pointed out in the scene of lexicography of EFL dictionaries in the country. It should not need to be pointed out, yet apparently it must: the phonetic transcription of the pronunciation of the word *during* for American English. Almost all the Japanese EFL dictionaries including *Taishukan's Genius English-Japanese Dictionary* (5th ed., henceforth abbreviated as *Genius*[5]), *The Wisdom English-Japanese Dictionary* (3rd ed., *Wisdom*[3]), and *O-Lex English-Japanese Dictionary* (2nd ed., *O-Lex*[2]) give /ˈdəːrɪŋ/ to the pronunciation of the word for American English.[1] We could ask how on earth Japanese EFL dictionaries have this type of pronunciation recommended and provided for *during* for American English given the fact that almost all the English (both American and British) dictionaries do not list this form for the word during for American pronunciation. The problem seems more controversial mainly because the word *during* is one of the basic words in English, which means it is most frequently used both in spoken and written English. And a puzzle arise: why this situation happened in the Japanese EFL dictionaries. This paper attempts to find an answer for the query from a descriptive linguistics viewpoint.

What Dictionaries Tell us about Pronunciation of *during* in American English: American and British Dictionaries

Our point of departure is to consider what phonetic information current dictionaries give about the pronunciation of the word *during*.

Table 1 below gives an outline of how *during* in American English is phonetically represented in British and American EFL dictionaries.

Table 1 Comparison between BrE and AmE pronuncations of *during* in different dictionaries

Comparison between BrE and AmE pronuncations of *during* in different dictionaries

		BrE	AmE
British Dictionaries	*LDOCE*[6] (2014, 2017 Online)	/ˈdjʊərɪŋ/	/ˈdʊrɪŋ/
	OALD[6] (2015, 2017 Online)	/ˈdjʊərɪŋ/	/ˈdʊrɪŋ/ (NAmE)
	Oxford Living Dictionaries North American English (Online)		/ˈd(j)ʊrɪŋ/ /ˈd(y)o͝oriNG/
	Oxford Living Dictionaries British & World English (Online)	/ˈdjʊərɪŋ/	
	COBUILD Advanced Learner's Dic (2014[8], 2017 Online)	/ˈdjʊərɪŋ/	/ˈdʊrɪŋ/
	CED[12] (2014)	/ˈdjʊərɪŋ/	
	CALD[4] (2013)	/ˈdjʊərɪŋ/	/ˈdʊrˌɪŋ/
	CEPD[18] (2011)	/ˈdʒʊəˌrɪŋ , ˈdʒɔː- , ˈdjʊə- , ˈdjɔː-/	/ˈdʊrˌɪŋ , ˈdjʊr- , ˈdɜː-/
American Dictionaries	*Webster's New World College Dictionary*[4] (2010, 2017 Online)		/ˈdʊrɪŋ; door'iŋ//ˈdjʊrɪŋ; dyoor'iŋ/ /ˈdɜrɪŋ; dʉr'iŋ/
	RHD[2] (1993, 2017 Online)		/ˈdʊrɪŋ; door'iŋ/ /ˈdjʊrɪŋ; dyoor'iŋ/
	Webster[3] (1981, 2017 Online)		/ˈdʊrɪŋ/ /ˈdjʊrɪŋ/ 'dûr-iŋ also 'dyûr-
	AHD[5] (2017 Online)		/ˈdʊrɪŋ/ /ˈdjʊrɪŋ/ (do͞or'ĭng, dyo͞or'-)
	OAD (2011)*		/ˈdʊrɪŋ/
	Merriam-Webster's Advanced Eng Dic (2008, 2017)	/ˈdjʊrɪŋ/	/ˈdʊrɪŋ/
	COBUILD Advanced American Eng Dic (2016[2] ♦♦♦)*		/ˈdʊərɪŋ/
	LAAD[3] (2013, 2017 Online)* ●●● S1 W1		/ˈdʊrɪŋ/

**OAD*, *COBUILD AAED*, and *LAAD*, published by British publishers, yet describing the American English, are grouped into the American Dictionaries.

What does this table mean for the phonetic transcription of *during* in American English? Interestingly, all the dictionaries consulted above but the *WNCD*[4] and *CEPD*[18] give /ˈd(j)ʊrɪŋ/ to *during* unlike the Japanese EFL dictionaries, which list /ˈdɜːrɪŋ/. [2]

John Wells's *LPD*

Before continuing, let us turn to John Wells's *Longman Pronunciation Dictionary* (*LPD*), one of the most reliable English pronunciation dictionaries, to see what the entry of *during* in *LPD* looks like.

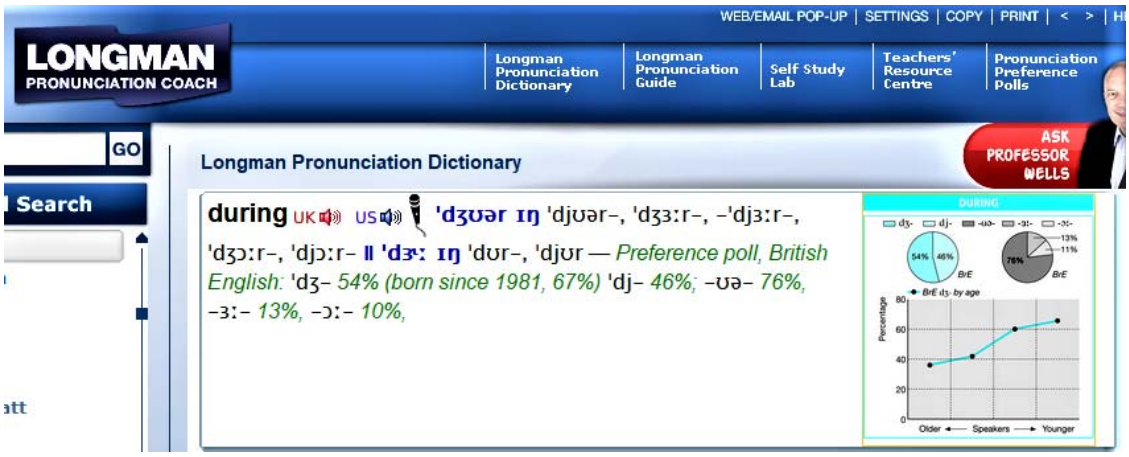


Figure 1 the entry *during* in *LPD*[3]

The entry of this sort suggests that Japanese EFL dictionaries just follow *LPD* as far as the pronunciation of *during* for American English is concerned. I shall come back to this problem just below.

We might ask whether this kind of transcription is appropriate for Japanese EFL dictionaries, or more broadly, English-Japanese dictionaries in general. Unfortunately my answer is in the negative. What follows discusses why my answer is no from a descriptive linguistics point of view.

The *LPD* since its launch in 1990 has been consistent with the variant /ˈdɝːɪŋ/ as the recommended pronunciation for the main variety of American English. A quick comparison between G[1] and G[2-5] clearly shows that the adoption of /ˈdɜːrɪŋ/ by G[2] and its later editions is based on and follows *LPD*[1] (1990), since G[1] gives /ˈd(j)ʊrɪŋ/. The first edition of *Genius English-Japanese Dictionary*, which I participated in compiling as a member of the editors, was published in 1988 when there was no *LPD* published and it was utterly impossible for the then lexicographers of Japan to refer to *LPD* to confirm and decide upon the recommended pronunciation for any word whatsoever.

Sadly and more importantly, these Japanese EFL dictionaries do not provide any evidence at all that supports a change in phonetic transcription of *during* from /ˈd(j)ʊrɪŋ/ to /ˈdɜːrɪŋ/ for American English. Therefore it may well be true that G[2] along with its later

[2] The transcription with an optional yod /ˈd(j)ʊrɪŋ/ means that Americans usually pronounce the word as /ˈdʊrɪŋ/ while Britons use a palatalised variety /ˈdjʊrɪŋ/. See Lindsey (2018) for a new approach to the British vowel system. He is attempting to describe the vowels of British English (BrE) as they are, rather than as they're implied to be by the most familiar dictionary symbols.

editions and other popular Japanese EFL dictionaries just follow the contents of the entry in *LPD* without checking by themselves.

Being a linguist rather than a phonetician, I am also a regular user of John Wells's reliable *Longman Pronunciation Dictionary* since the first edition. I had just noticed a small point in the current and past editions of *LPD*.

The current edition as well as the past two editions gives /ˈdɝːɪŋ/ as the recommended pronunciation of *during* for American English (see Figure 1 for the entry in the current *LPD*).

I was just wondering if he has any reason or evidence for listing this pronunciation as the suggested one. As far as I see, I could not find this pronunciation as the suggested one in the current American dictionaries such as *AHD*, *RHD*, *Merriam-Webster's*, etc. They seem to go for the more conservative /ʻdʊr-, ˈdjʊr-/.

I was interested to know why LPD has the form /ˈdɝːɪŋ/ for the main American English pronunciation for *during* in contrast to almost all the EFL dictionaries published both sides of the Atlantic Ocean which represent otherwise, I sent Professor John Wells an email asking the reason. [3]

In his reply to my inquiry, Professor Wells says:

> You are quite right to query this. Clearly, the main AmE pronunciation given in *LPD* is wrong. It should be /ˈdʊr ɪŋ/.
> It is thirty years since I worked on the *D* section of the first edition, so I certainly cannot now remember how or why I came to write the entry as I did.
> Thanks for drawing it to my attention.

Now that Professor Wells admits that the entry of *during* in his *LPD* is wrong and to be corrected, it is clear that the recommended pronunciation of *during* for American English IS /ˈdʊr ɪŋ/ rather than /ˈdɝːɪŋ/.

The Real Facts Offered by Forvo.com

I took a look at Forvo.com (https://ja.forvo.com/word/during/#en)[4], which offers registered users of native speaker of English the opportunity to record themselves pronouncing words or phrases. According to my survey of pronunciation of *during* on this site, I hear only four out of 10 North Americans (including two Canadians) who posted their pronunciations of *during* actually use the pronunciation /ˈdɝːɪŋ/:[5]

/ˈdɝːɪŋ/ SeanMauch, Tong, migueld, Atalina

The rest of the results is given below: one American gives /ˈdʊrɪŋ/ while /ˈdjʊrɪŋ/ is produced by four Britons, three Americans, and two Canadians.

/ˈdʊrɪŋ/ mmdills22

[3] Dispatched in October 2017.
[4] Accessed on 17th May 2018
[5] Each form of pronunciation is followed by the user name of a speaker who recorded that form.

/ˈdjʊriŋ/ TopQuark, BritishEnglish, mstormw, onelonegypsy, Slick, itiwat, kstone11, *IAmMaidOfTheMist*, *NatalieHosge*[6]

This result further confirms that an American variety /ˈdɝːɪŋ/ is not so wide-spread. It also shows that as many as five North Americans including two Canadians really use a more conservative (or British-accented) /ˈdjʊriŋ/.

Nursing the *CURE* Vowel

To examine further a choice between /ˈdəːrɪŋ/ and /ˈd(j)ʊriŋ/, I put a question to the English Language & Usage Stack Exchange on the internet, which is a question and answer site for linguists, etymologists, and serious English language enthusiasts to collect further information about a choice between /ˈdəːrɪŋ/ and /ˈd(j)ʊriŋ/. [7] I have received a variety of responses from the registered contributors. All in all, what the traditional view says is that the pronunciations /ˈdəːrɪŋ/ and /ˈdʊriŋ/ are both used by lots of Americans. If we use either one of these, nobody is going to think we are speaking strangely. In some parts of the country (i.e. USA or Canada), the vowel combination /ʊr/ is slowly disappearing from English and being variously replaced by /əːr/, /ɔːr/, and /uːər/, depending on the exact word and the part of the country he or she lives in. And if we pronounce *during* /ˈdjʊriŋ/, which does not seem to be very common anymore in American English, nobody will be going to notice that, either. That form of pronunciation will allow them to think we have learned English from a British teacher.

One of the contributors to this site mentions the *Nurse-Cure* merger and thinks our vowel change with *during* it as an instance of this type of phonetic change. This merger, by the way, occurs in England as well as in USA. In East Anglia, England, a ***cure–nurse* merger** in which words like *fury* merge to the sound of *furry* [ɜː] is common, especially after palatal and palatoalveolar consonants, so that *sure* is often pronounced [ʃɜː] (which is also a common single-word merger in American English, in which the word *sure* is often /ʃɚ/); yod-dropping may apply as well, yielding pronunciations such as [pɜː] for *pure*. Other pronunciations in *cure–fir* merging dialects include /pjɚ/ *pure*, /ˈk(j)ɚiəs/ *curious*, /ˈb(j)ɚoʊ / *bureau*, /ˈm(j)ɚəl/ *mural* (cf. *Wikipedia* (s.v. English-language vowel changes before historic /r/).

The same contributor says:

> The pronunciation of *during* with the NURSE vowel /ɜr/ rather than the CURE vowel /ʊr/ is an example of a current trend in rhotic American accents of a merger — called, appropriately enough, the NURSE-CURE merger — affecting some words in the CURE group, especially among younger speakers.
> Dictionaries have generally been slow to acknowledge this change. While
>
> *Merriam-Webster* lists the pronunciation of *during* as:
>
> /ˈdu̇r-iŋ also ˈdyu̇r-/ = ('dʊr-iŋ, djʊr-)

[6] User names in romans indicate Americans, those underlined Britons, those in italics Canadians
[7] https://english.stackexchange.com/questions/, posted on 6th May 2018; for the whole transaction, go to https://english.stackexchange.com/questions/444965/pronunciation-of-during-in-na-english?noredirect=1#comment1072638_444965.

the entry for *cure* itself offers the NURSE pronunciation as an alternate:

/ˈkyu̇r , ˈkyər/ = ('kjʊr, 'kjɜr)

For both these examples I have supplied a more standard IPA transcription to the right. Your dictionaries in Japan, then, seem to be ahead of the curve in recording this pronunciation. This also suggests that CURE may soon lose its usefulness as a member of a lexical set.

Now I would suggest that the cause of this merger is the result of the slight lip rounding of an American r reducing the fuller rounding of the /ʊ/, yielding a vowel actually closer to a /ɯ/. That vowel, however, is only found as one possibility in the New Zealand treacLE set, so American ears are more likely to hear a schwa or an /ɜ/, if they prefer not having a schwa in an accented syllable. But if there is the slightest lip rounding the vowel cannot be /ə/ or /ɜ/.

The difference between the two pronunciations is subtle and not likely noticed by native speakers, especially because a preposition rarely receives sentence stress.

As is claimed by the contributor, dictionaries are very slow to register this sort of merger in vowels and this merger as is seen with *during* is not recorded with *pure* and all of its homophonous words in EFL dictionaries both in Japan and USA.

It seems to me that this phonetic mechanism certainly explains what happens with *during* in North America and it is also quite reasonable to reckon it as an instance of more general 'vowel change before /r/' in rhotic dialects. However, this explanation unfortunately does not say much about how much and how frequently this phenomena spreads in North America.

The survey so far confirms that it is wrong to recommend /ˈdəːrɪŋ/ as the main American English pronunciation for *during* despite the fact that most Japanese EFL dictionaries actually do quite wrongly. And in doing so, they apparently ignore the actual phonetic facts, which will not or cannot be sufficiently accessed without carrying out a large-scale research into the phonetic data.

Conclusion

It is needless to say that dictionaries for language learning should provide information about the language but they should also help learners to learn the language. In the light of the latter point, what lexicographers provide as the recommended variety of pronunciation for a given word (or lexical item) is significantly important both for the learners and the lexicographers alike.

Generalisations on the basis of a limited number of dictionaries as has been wrongly done with Japanese EFL dictionaries, should be carefully taken care of.

In the sections above, I have argue that what lexicographers bear in mind is that they are required to describe and represent the phonetic data as it is based on their or others' large-scale research into the pronunciation varieties rather than just making generalisations through reading available literatures or dictionaries. This applies not just to the lexicographers but to phoneticians in general as well. It is also an attitude that EFL teachers should continue to accept as their obligation.

From the arguments above, the following conclusions can be drawn.

- Phonetic transcriptions in EFL in general should reflect the real pronunciations of the relevant lexical items.
- Check with informants whether a given pronunciation really reflects how the

word is actually pronounced.
- Collect as much information as is required from large-scale corpora to decide the most recommendable variety of pronunciation.

Description of language including the phonetic transcription of English should be presented (to non-native school children, university students, and the general public) as exactly as the English language appears to be. Just go back to the basic principle of descriptive linguistics.

References

English-language vowel changes before historic /r/. (n.d.). In *Wikipedia*. Retrieved May 18, 2018, from https://en.wikipedia.org/wiki/English-language_vowel_changes_before_historic_/r/#Cure%E2%80%93nurse_merger

Lindsey, G. (2018, May 18). The British English Vowel System. Retrieved from http://englishspeechservices.com/blog/british-vowels/

English-Japanese EFL Dictionaries

Taishukan's Genius English-Japanese Dictionary. (1994[2], 2002[3], 2006[4], 2014[5]). Tokyo, Japan: Taishukan.

O-Lex English-Japanese Dictionary. (2008, 2013[2]). Tokyo, Japan: Obunsha Publishers.

The Wisdom English-Japanese Dictionary. (2003, 2006[2], 2012[3]). Tokyo, Japan: Sanseido Publishers.

British Dictionaries

Cambridge Advanced Learner's Dictionary. (2005[2], 2008[3], 2013[4]). Cambridge, England: Cambridge University Press. [*CALD*]

Cambridge English Pronouncing Dictionary. (2003[16], 2006[17], 2011[18]). Cambridge, England: Cambridge University Press. [*CEPD*]

Collins Cobuild Advanced Learner's Dictionary (Collins Cobuild English Language Dictionary (Advanced Learner's English Dictionary). (1987, 1995[2], 2001[3], 2003[4], 2006[5], 2009[6], 2012[7], 2014[8]). Glasgow, Scotland: HarperCollins Publishers. [*COBUILD Advanced Learner's Dic*]

Collins English Dictionary: Complete and Unabridged. (2014[12]). Glasgow, Scotland: HarperCollins Publishers. [*CED*]

Longman Dictionary of Contemporary English. (2009[5], 2014[6]). Harlow, England: Pearson. [*LDOCE*]

Longman Pronunciation Dictionary. (2000[2], 2008[3]). Harlow, England: Pearson. [*LPD*]

Oxford Advanced Learner's Dictionary. (2010[8], 2015[9]). Oxford, England: Oxford University Press. [*OALD*]

Oxford Living Dictionaries British & World English (Online) . Oxford, England: Oxford University Press. Available from https://en.oxforddictionaries.com/english

American Dictionaries

The American Heritage Dictionary of the English Language. (2011[5]). Boston, MA: Houghton Mifflin Harcourt. [*AHD*]

Collins Cobuild Advanced Dictionary of American English. (2007, 2016[2]). Glasgow, Scotland: HarperCollins Publishers. [*COBUILD Advanced American Eng Dic*]

Longman Advanced American Dictionary. (2000, 2007, 2013[3]). Harlow, England: Pearson. [*LAAD*]

Merriam-Webster's Advanced Learner's English Dictionary. (2008, 2017[2]). Springfield, MA: Merriam-Webster,Inc. [*Merriam-Webster's Advanced Eng Dic*]

Oxford American Dictionary. (2011). Oxford, England: Oxford University Press. [*OAD*]

Oxford Living Dictionaries North American English (Online) . Oxford, England: Oxford University Press. Available from https://en.oxforddictionaries.com/english

The Random House Unabridged Dictionary. (1993[2], 2013 Online). New York, NY: Random House Reference. [*RHD*]

Webster's New World College Dictionary. (2010[4], 2017 Online). Boston, MA: Houghton Mifflin Harcourt. Available from http://websters.yourdictionary.com/

Webster's Third New International Dictionary International Dictionary: Since 1847 the Ultimate Word Authority for Schools, Libraries, Courts, Homes, and Offices. (2002). Springfield, MA: Merriam-Webster, Inc. [*Merriam-Webster's*]

Gender Orientation in Lexis, Corpora and Dictionaries

Lan LI & Yue GU
The Chinese University of Hong Kong, Shenzhen
lanli@cuhk.edu.cn

Abstract

The topic of inclusive language has long been marginalized in lexical studies, partially because of the notion of political correctness, which remains a bitter controversy. Understanding gender orientation in a language distinctively specifies social background into linguistic contexts and can also demonstrate the cultural difference between different tongues. Besides all the regulations and debates, language is not a sheer mirror of the society, but itself is preserved as a unique institution by people. This paper aims to explore a symbiosis between people and language by transforming the invisible social structure into traceable linguistic records. Through morphological and corpus approaches, it investigates gender-marked compounds and gender-specific references used in the English language. Their counterparts in bilingual dictionaries are also discussed. Discussion of sexism in lexical items and their use can illustrate the significance of inclusive language in academic as well as everyday communication.

Keywords: gender; compounds; inclusive language; sexist language

Introduction

Inclusive language, according to Collin's Dictionary, is 'language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as "*man*", "*mankind*". Using inclusive language is another way to practice political correctness.

The study of gender and communication has largely been led by feminist scholars. Such scholarship has tended to focus on power imbalances and the way in which these are created and maintained by discourse. The general approach of early researchers took a dominance perspective: i.e. men used language as a means to purposely suppress women (Lakoff, 1975; Spender, 1981). Women were seen as restricted in how they could use language (i.e. indirect and polite), and language itself contained sexualized bias against women. This study will focus on the latter

Different from languages like French, German, and Latin with full grammatical gender, modern English only expresses this specific form of the noun-class system through the third person singular personal pronouns and their possessive forms (Geville, 1994). It means, except for certain borrowed words from other languages, few of English nouns have masculine, feminine, or neuter classes.

However, this lack of gender agreement between nouns and modifiers doesn't guarantee complete gender neutrality. Morphemes, including affixes and roots, sometimes can carry a presumed gender notion. For instance, suffixes *–or* and *–ess* in words such as *actor* and *actress*, actually reveal straightforward masculinity and femininity. Conventionally, the word or the stem *man* (pl. *men*) can be independently used to represent all human beings in a broad sense, as in the Declaration of Independence of America written by Thomas Jefferson in 1776.

> *We hold these truths to be self-evident, that all **men** are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.*
>
> (Jefferson, 1776)

The lexeme *man* can also be used as a free root, generalising both males and females (e.g. *mankind*, *manpower*, and *man of letters*). As time passes by, when it comes to an individual of unmentioned gender, the prior option in personal pronouns or their adjectives is always the masculine. Sentences like the following prevail in the daily life:

> *If a doctor wants to announce the diagnosis of a terminal disease, **he** will first inform the relatives instead of turning to the patient directly.*

It is generally accepted that language is far more than a communication medium. It is an interpersonal activity and a matter of social practice. And, it is not neutral. In practice, however, people could deliver discrimination either by using specific lexemes or by using them in a particular sense, while we can advocate inclusivity through modest changes in the wording.

This paper aims to explore a symbiosis between people and language by transforming the invisible social structure into traceable linguistic records. By comparing the context and frequency of two symbolic morphemes in solid compounds — *man* and *woman*, it traces the gender orientation in English vocabulary, particularly the missing narration for women. It will also examine a gender-neutral lexical morpheme — *person*, as a typical substitute for distinctive gender roots in inclusive language.

Methodology

The project takes a data-driven approach combined with dictionary exploration. The data employed is The Corpus of Contemporary American English (COCA https://corpus.byu.edu/coca/), the largest freely-available online corpus of American English, compiled by M. Davis at Brigham Young University. It is composed of over 560 million words classified into spoken, fiction, journal, and academic texts from 1990 to 2017. [1] Davies (2011) regards it as the most widely-used corpora at present. COCA was academically preferred for its massive database, intensive categories, free accessibility and persuasive timeliness.

The first stage is to download the lists of search results of three strings:**man*, **woman* and **person*. The data was then cleaned to delete proper nouns such as personal names and nationalities. The data retrieval yielded sufficient information for the second step- analysis and comparison. The third step is to have a couple of case studies, looking into typical examples in context to interpret the use of inclusive-language.

Findings and discussion

1. Frequency of use

The frequency of *man-words* is 1,214,151, taking up 1.86% of the total words in the COCA corpus, while compounds with *–woman* have a much lower frequency, only one-sixth of the *man*-words, amounting to 213,611. Exploring the compounds, we found that about 40% of *man*-words are proper nouns, such as family names *Whiteman, Bowman, Goldman, Goodman*; nationalities such as *Englishman, Irishman.* The occurrences of *person*-words are slightly lower than that of *woman*-words. Table 1 shows the summary of the gender-related words.

Table 1 Frequency of words with *man, woman* and *person*

Search string	Raw frequency	Frequency of compound	Number of unique word
**man*	1,214,151	844,549	8,839
**woman*	213,611	15,625	480
**person*	155,701	10,712	582

From the statistics, we may tell that English is not a gender-neutral language. Man-words have an obvious dominance; many of them have been recorded in dictionaries. Some of the compounds with *woman* and *person* are rarely used that their qualification as a word remains to be tested. It can be assumed that gender-specific words have been invented and used by individual writers therefore many of them (e.g. *outdoorswoman*, *boogiewoman*) have not been recorded in authoritative dictionaries such as Oxford English Dictionaries and even fewer can be found in Big Five learner's dictionaries. Bilingual dictionaries, such as New English-Chinese Dictionary, do not include many gender-specific words probably because gender is used periphrastic in Chinese by adding woman adjective before nouns, as in 女法官 (woman lawyer)，女教师 (woman teacher) and 女运动员 (sportswoman), rather than as a component of words.

[1] Extracted from documents on COCA (Last reviewed:13th December 2017) https://corpus.byu.edu/coca/help/texts.asp/

2. Semantic field

Of all the *man*-words and *woman*-words, our interest is to find out differences between professions. Some jobs were normally done by men in the past, and their names had no form for women (e.g. *fireman, fisherman*). Such practice seems to have had a strong influence on the English vocabulary and generated many gender-biased words such as *congresswoman, spaceman, deliveryman, gunnman, chapman, letterman* and *lineman*. This could explain why there are much more *man*-words than *woman*-words. With the time change, man and woman can do almost everything, so it is not correct to use *man*-words which exclude woman in many professions. To make the research manageable, we examined 600 man-words and 480 woman-words and classified them into seven semantic fields. *Man*-words of English names and company names were excluded. In Table 2 and 3, 'Type' represents unique words in the field, and 'Frequency' refers to the words' additive occurrences in the whole corpus (see Table 2 & 3).

Table 2. Classification of "*man*" compounds

	Type	Frequency	Example
Profession	100	127,709	*statesman, foreman, chapman, doorman, fisherman*
Character	15	30,005	*ironman, madman, strongman, wiseman*
Race	19	6,397	*Irishman, Frenchman, blackman*
Fiction	6	6841	*superman, batman, wolfman, sandman*
Sports	13	11,586	*bowman, lineman, baseman, defenseman*
Community	8	1326	*wingman, kinsman, layman*
Object	5	2243	*snowman, walkman, talisman*

Table 3. Classification of "*woman*" compounds

	Type	Frequency	Example
Profession	57	14,859	*congresswoman, assemblywoman anchorwoman, forewoman newswoman, clergywoman birdwoman, craftswoman*
Character	11	424	*superwoman, catwoman*
Race	8	359	*Englishwoman, Scotswoman*
Fiction	5	796	*spiderwoman, weyrwoman*
Sports	4	94	*sportswoman, defensewoman*

Community	4	7	*wingwoman, kinswoman countrywoman, laywoman*

Notably, both "*man*" and "*woman*" compounds contain a large number of words relating to professions and crafts, but the number of professional terms of "*man*-words" is almost twice as many as "*woman*-words". Women's narration seems to be neglected in the fields of sports and faith communities. Women are more likely to be portrayed in fictions than in sports activities and social affairs. On the whole, although parallel terms exist in each semantic field above, a great disparity does stand out regarding variety and frequency.

Case studies

After a brief quantitative analysis of the present situation of words used in inclusive language, this section will discuss a couple of prominent gender-specific pairs from three aspects to examine the practical effects these linguistic shifts bring.

1. Semantic prosody: *wiseman* & *wisewoman*

Semantic prosody refers to positive or negative context words belongs to. *Madman* and *madwoman* both are negative describing someone who behaves in a wild, uncontrolled way. The male-female pair with the root *wise*, according to Oxford English Dictionary, can mean entirely different. While the original word *wiseman* solely praises a person's intelligence, its variation *wisewoman* has two different denotations. One means *witch*; another refers to *the midwife*. Both of them constrain sound possibility for women to a perioheral group — women who engage in particular vocations (often being viewed as trivial matters). It may imply a message: Wisdom is rare for women and women with wit are estranged from the mainstream. Thus, despite the formal diversity, discriminatory concerns remain.

2. The use of *policeman* and *policewoman*

Among the pairs of male-female compounds, *policeman* and *policewoman* are probably the ones mostly explored. So it is vital to confirm how the two are used in real-life communication. Google Ngram Viewer[2] (Figure 1) reveals the trend of the use of gender-specific *policeman* and *policewoman*, and also the gender inclusive compound *police officer*. The findings show a striking gap between the supposed-to-be equivalents with an obtrusive widening. Problems arise when people invent new words while not using them, not only resulting in futility but also misleading people about the continuous linguistic issue. The growing trend is to use *police officer*. A gender-neutral word *policeperson* was also invented but has rarely been used, with only three occurrences in the 560 million-words COCA. Similar trend can be found in many other groups, for instance, *layman, laywoman layperson*; *anchorman, anchorwoman, anchorperson.*

[2] The Google Books Ngram Viewer provides upgraded inquiries into the usage of phrases based on datasets from book scanning.
https://books.google.com/ngrams

Figure 1. The comparison of *policeman, policewoman and police officer* on Google Ngram Viewer

Woman-words and person-words were mostly developed from open compounds to solid compounds. They are used by sensitive users to avoid exclusion of woman. The chart shows a clear trend that the use of gender-biased term *policeman* has dropped greatly since the 1970s and the gender-neutral word *police officer* has kept increase in the last thirty years. This is an indication that more and more writers are aware of the role of inclusive language in the contemporary society.

3. *Person* as a stem of gender-neutral compound

As the society has become more open and pluralistic, people identify others with a more varied gender spectrum instead of the male-female binary. In an attempt to avoid exclusivity, words seek ways to present communities in shared features. People are first and foremost humans, with a more generic hypernym rather than categorised hyponyms; unnecessary division or tendentiousness could be avoided. Therefore, "person-centred" language becomes a neutral substitute for *man* and *woman*. Salesman and saleswoman could both be addressed as a *salesperson*, chairman or chairwoman could be written to the *chairperson* or simply, *chair*, and a fireman has reformed to *a firefighter*, etc. The idea is appealing, but the practice is yet to be facilitated. In COCA, words with *-person* occur 10,712 times. There are 582 unique words, 302 of which are bound with numbers either as a close compound or a hyphenated compound, for example, *twoperson, 100-person*, and *millionperson*. They were excluded from the list. Table 4 lists the *person* compounds with content words relating to the particular profession, craft, character, or region.

Table 4. "*-person*" compounds in COCA

Item	Frequency	Item	Frequency	Item	Frequency
Spokesperson	2860	Policeperson	3	Stockperson	1
Salesperson	667	Snowperson	3	Strongperson	1
Chairperson	560	Deliveryperson	3	Townperson	1
Layperson	560	Counterperson	3	Pointperson	1

Businessperson	139	Headperson	3	Smartperson	1
Congressperson	72	Fisherperson	3	Showperson	1
Townsperson	64	Midshipperson	2	Significantperson	1
Craftsperson	47	Henchperson	2	Selectperson	1
Clergyperson	25	Cowperson	2	Serviceperson	1
Anchorperson	19	Everyperson	2	Fireperson	1
Newsperson	19	Spokeperson	2	Dutyperson	1
Sportsperson	15	Spaceperson	2	Disabledperson	1
Serviceperson	13	Pitchperson	2	Deadperson	1
Councilperson	11	Woodsperson	2	Countryperson	1
Waitperson	10	Batperson	2	Houseperson	1
Draftsperson	8	Committeeperson	2	Helmsperson	1
Cameraperson	7	Busperson	2	Gentleperson	1
Statesperson	7	Caveperson	2	Gingerperson	1
Weatherperson	7	Cattleperson	1	Fishperson	1
Repairperson	7	Businesssperson	1	Freethinkingperson	1
Outdoorsperson	7	Bellperson	1	Frenchperson	1
Crewperson	6	Birdperson	1	Ironperson	1
Tradesperson	6	Blackperson	1	Lazyperson	1
Frontperson	5	Bondperson	1	Letterperson	1
Mailperson	4	Boogyperson	1	Lineperson	1
Missingperson	4	Walkperson	1	Medicinepers on	1
Handyperson	4	Watchperson	1	Marksperson	1
Staffperson	4	Waterperson	1	Layperson	1
Stuntperson	4	Salesperson	1	Journeyperson	1

The	Cochairperson	4	Stickperson	1	Nurseryperson	1	data
					Overweightperson	1	

indicates a frequency difference of *person*-words, similar to "*woman"* compounds (see Appendix), person compounds represent limited careers, such as politics, commerce and crafts. There are many guidelines for non-descriminatory writings, suggesting substitutions to preclude sexist language, for example, use *scholar* and *academic* instead of *man of letters*; use *personnel, staff,* or *human resources* to replace *manpower*.

Conclusion

Language is an indicator of social norms speakers place on others and themselves. Even the modern society sustains concealed patriarchies, morphological change in the English vocabulary has more or less reflected the increased awareness of gender-neutral language. Discourse is shaped with the participation of ideologies and institutions, and communication, especially pragmatic conversations can create and construct our cultural environment. The increasing cultural diversity requires people to recognise and unveil the imparities in the linguistic repertoire. Inclusive language went on stage in the late 20th century, with the aim of offsetting linguistic imbalance and crossing unconscious language barriers. All in all, being aware of the fact that the evolution of language lies not only in word formation but also in the interpretation and usage of words, it is necessary to avoid sexist language in order not to offend others.

The concept of inclusive language is important to language learners and users. As Nielsen points out, "students must receive adequate instruction on avoiding sexist language, particularly in textbooks intended to help students develop and refine their skills" (Nielsen 1998: 56). We have no intention to advocate more gender natural words, but wish to raise a question: Is it necessary to mention one's gender when a professional is referred to? A commitment to inclusive language is an important attribute of a modern, diverse and inclusive society.

Reference List

Corbett, G. (1994). Gender and gender systems. En R. Asher (ed.). *The encyclopedia of language and linguistics* (pp.1347-1353). Oxford: Pergamon Press.

Curzan, A. (2003) *Gender shifts in the history of English.* Cambridge: Cambridge University Press.

Davies, M. (2014). Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics, 19*(3), 401-16.

Davies, M. (2011). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25, 447-65.

Fischer, R. (2002). Words, meaning and vocabulary: An introduction to modern English lexicology. *AAA: Arbeiten Aus Anglistik Und Amerikanistik, 27*(1), 87-89. Retrieved from http://www.jstor.org/stable/43025659/

Lakoff, R. (1975). *Language and woman's place.* New York: Harper and Row.

Nielsen, E. (1998). Linguistic sexism in business writing textbooks. *Journal of Advanced Composition, 8*(1), 55-65.

Prewitt-Freilino, J., T. Andrew Caswell, T. and Laakso, E. (2012). The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66, 268–281. DOI 10.1007/s11199-011-0083-5

Soames, S. (2010). *Philosophy of language*. Princeton | Oxford: Princeton University Press.

Spender, D. (1980). *Man made language.* London: Routledge & Kegan Paul.
Thomas J. (1776). Declaration of Independence. In Congress, July 4, 1776, a Declaration of Independence by the Representatives of the United States of America, in General Congress Assembley. World Digital Library. Retrieved from: https://www.wdl.org/en/
Wright, L. (2017). Inclusive language guidelines (2nd Edition). Retrieved from https://www.uow.edu.au/about/policy/UOW140611.html/

Appendix 1 ***"Man"* compounds in COCA**

Context	Frequency	Context	Frequency	Context	Frequency
Chairman	37000	Patrolman	573	Lawman	254
Spokesman	17512	Hillman	555	Guardsman	246
Congressman	16049	Horseman	541	Chinaman	243
Freshman	8858	Layman	541	Birdman	238
Gentleman	7494	Newsman	470	Serviceman	235
Businessman	6094	Clergyman	467	Infantryman	233
Freeman	4650	Nobleman	463	Oilman	227
Salesman	4343	Weatherman	443	Newspaperman	227
Policeman	4135	Pitman	442	Sandman	226
Goodman	3714	Waterman	436	Bogeyman	224
Batman	3625	Talisman	424	Helmsman	221
Baseman	3230	Everyman	420	Kingman	220
Letterman	2874	Showman	414	Minuteman	218
Gunman	2791	Airman	407	Spiderman	210
Fisherman	2737	Iceman	407	Woodman	208
Bowman	2320	Walkman	407	Boatman	208
Superman	2315	Strongman	402	Marksman	204
Lineman	1665	Coachman	393	Draftsman	202
Statesman	1567	Repairman	385	Henchman	192
Councilman	1528	Journeyman	385	Swingman	187
Cameraman	1292	Spearman	374	Landman	184
Craftsman	1234	Stockman	373	Landsman	181
Freedman	1175	Blackman	363	Bellman	180
Frenchman	1139	Hackman	354	Pitchman	177
Madman	1133	Footman	336	Bondsman	171
Englishman	1120	Hangman	326	Bookman	171
Fireman	1117	Ironman	323	Brinkman	171
Doorman	937	Waterman	436	Englishwoman	167
Pullman	900	Dutchman	323	Rifleman	167

Workman	876	Spearman	374	Deliveryman	157
Seaman	845	Stockman	373	Bluesman	157
Defenseman	827	Blackman	363	Klansman	155
Assemblyman	714	Hackman	354	Markman	154
Mailman	708	Footman	336	Backman	154
Huntsman	700	Hangman	326	Bushman	149
Sportsman	647	Ironman	323	Wolfman	142
Postman	594	Waterman	436	Spaceman	140
Glassman	577	Dutchman	323	Wingman	140
Watchman	537	Shipman	317	Houseman	139
Wiseman	534	Crewman	303	Oklahoman	137
Irishman	531	Caveman	302	Corpsman	131
Snowman	512	Pressman	294	Stuntman	126
Middleman	512	Hitman	291	Woolman	126
Fishman	505	Carman	288	Needleman	125
Anchorman	504	Scotsman	286	Midshipman	123
Handyman	491	Woodsman	284	Cattleman	122
Stillman	474	Countryman	279	Workingman	119
Newsman	470	Barman	275	Herdsman	114
Clergyman	467	Headman	269	Ferryman	111
Nobleman	463	Wellman	263	Kirkman	106
Weatherman	443	Outdoorsman	263		
Pitman	442	Milkman	258		

Appendix 2. "*Woman*" Compounds in COCA

Context	Frequency	Context	Frequency	Context	Frequency
Spokeswoman	6221	Headwoman	17	Townswoman	5
Congresswoman	2272	Newspaperwoman	15	Henchwoman	4
Chairwoman	1114	Stateswoman	14	Needlewoman	4
Businesswoman	681	Camerawoman	13	Snowwoman	4
Catwoman	595	Servicewoman	13	Swordswoman	4
Councilwoman	483	Wingwoman	12	Wisewoman	4
Saleswoman	390	Birdwoman	11	Wildwoman	4
Policewoman	207	Craftswoman	11	Whitewoman	3
Englishwoman	167	Clergywoman	10	Repairwoman	3
Superwoman	164	Tribeswoman	10	Upperclasswoman	3
Madwoman	148	Spiderwoman	9	Ironwoman	3
Assemblywoman	137	Churchwoman	9	Doorwoman	3
Anchorwoman	128	Selectwoman	8	Herbwoman	3
Noblewoman	101	Plantswoman	8	Nurserywoman	3
Frenchwoman	89	Washwoman	7	Militiawoman	3
Washerwoman	72	Clubwoman	7	Laundrywoman	3
Newswoman	61	Cavewoman	7	Boogiewoman	3

Gentlewoman	60	Bondswoman	6	Brakewoman	3
Everywoman	57	Crewwoman	6	Bagwoman	3
Horsewoman	52	Markswoman	6	Adwoman	2
Countrywoman	35	Outdoorswoman	6	Airwoman	2
Laywoman	35	Pitchwoman	6	Bondwoman	2
Sportswoman	32	Scotswoman	6	Boatwoman	2
Patrolwoman	29	Woodswoman	6	Middleman	2
Batwoman	25	Weatherwoman	5	Hitwoman	2
Stuntwoman	24	Freedwoman	5	Infantrywoman	2
Committeewoman	21	Handywoman	5	Spacewoman	2
Fisherwoman	21	Blackwoman	5	Career-woman	2

A Hypergraph Data Model for Building Multilingual Dictionary Applications

Louis Lecailliez
NLP Centre, Masaryk University
Botanická 68a, Brno, Czech Republic
louis.lecailliez@outlook.fr

Mathieu Mangeot
Université Savoie Mont Blanc, LIG
38000 Grenoble, France
mathieu.mangeot@imag.fr

Abstract

A non-negligible part of learners of an East Asian language have an interest for another tongue from East Asia that may share some common areal features such as the use of Chinese characters as well as limited word morphology. It makes sense to build for this niche a dictionary application that provides multiple languages in one bundle and allows easy navigation between them and in the lexicon. This paper describes a data model and a generic dictionary application architecture that addresses this use case.

The task of merging lexical resources with vastly differing micro-structures is complex. Even more so when updating it to include new data types or languages after release. In this regard, lexical networks are appealing: they solve the problem by exploding the micro-structure into data nodes and explicitly linking them with edges that can be discovered and traversed automatically. One of these approaches, The Linked Data, is gaining traction in lexicography. It is however plagued with issues within the Resource Description Framework (RDF) that backs it up. Most notably, the lack of three-valent relationships makes it harder than it should to handle the Chinese writing system.

We therefore came up with a simple and consistent hypergraph data model whose main features are: hyperlinks (links of arity greater than two), a flat type system (non-ontological lexical network) and annotations for both node and link instances. We propose a generic software architecture based on this model and illustrate it with a working mobile application. The user interface is constructed from independent components, allowing the display of complex data while increasing further its updatability and maintainability. We use data from the Revised Mandarin Chinese Dictionary of the Ministry of Education of Taiwan and augmented it with open-data Japanese readings to feed the prototype.

1. Introduction

It is quite common for students majoring in East Asian Studies to have an interest for another East Asian language: some universities actually provide double major degrees to meet this demand. A lot of vocabularies of Sinitic and Sinoxenic (Hashimoto, 1973) languages come from a common ancestry or were borrowed from each other, so they are related in meaning and pronunciations. It then makes sense to present learners with similar words of other languages he or she is learning while browsing an entry in a dictionary.

On the application developer side, the task of integrating different dictionaries for a given language is not easy. Creating a multilingual dictionary is even harder: each one has its own microstructure that doesn't necessarily play well with others. Moreover, the transposition from paper dictionaries to electronic ones led to digital resources (XML or databases) which retain a structure calqued on what is displayed to the end user (Polguère, 2012). Hence, the abstraction level needed to easily merge and use these resources is not provided.

Graph-based Data Models

Some electronic dictionary systems such as the one described by Mangeot *et al.* (2001) exhibit a latent graph structure while still being mostly based on dictionaries structured in a traditional fashion. Nonetheless, the groundbreaking work that made researchers start to rethink the modeling of dictionaries themselves is WordNet (Miller, 1995). Other lexical networks of two flavors were created: ontological and non-ontological (Polguère, 2014a). Amongst the former are the Semantic Web and the Linked Data frameworks which are used in past and ongoing research projects (Declerck, 2015). However, because the Semantic Web relies on Internet connectivity to reach its full potential, it is by definition not suitable for offline dictionaries, even if its technology may be used in an embedded way. In addition, theses frameworks are based on RDF which have their own share of problems.

This situation motivates us to create a data model that tries to capitalize on the essence of what makes the Semantic Web a potentially powerful technology — its underlying graph model — while not being limited in our implementation by the inherent complexity of RDF-based frameworks like Lemon (McCrae et al., 2011) in which support of East-Asian languages is lacking (Lecailliez, 2017).

2. Issues with Existing Data Models & Frameworks

Issues with the Relation Model

The major paradigm for organizing software data is the relational model. It is generally used in conjunction with a software that is layered in three main parts: the data layer which communicates with the database system to retrieve or store data, the business layer which contains the core logic of the software and the presentation layer that displays content to the end-user. Dictionary software are typically implemented in this way.

The main issue with the three-layered software on top of a relational database approach is that each modification of the schema requires a database migration, an update of the business logic and changes in the presentation layer. That is, every layer of the application is impacted.

Thus, we need a model which allows a dictionary to change its micro-structure frequently with the least possible impact to any other software layer. The model described in section 3 has the following properties: a change in schema would (1) not affect the business layer. Changes in data access (2) and presentation (3) layers made to support new data types are made by adding self-contained components; existing code does not require modifications.

Issues with RDF

While RDF data forms a graph (Powers, 2003), it is not a general-purpose graph modelling tool. The two central concepts of RDF are nodes and triples. Nodes exist in three flavors: resource, literal and blank. A triple (subject, property, object) denotes a "property" link between two nodes. The first major issue is that literals, which contains the actual data useful to a human being cannot be the subject of a relationship. The second issue is related to blank nodes, which are used to aggregate data from more than two nodes.

The Lemon Cookbook (McCrae *et al.*, 2010) gives a good example of these issues. In the Figure 1, a blank node is figured by square brackets and literals are double quoted. It encodes that some abstract resource has the written representation ⟨日本語⟩ and two "transliterations" ⟨にほんご⟩ and ⟨nihongo⟩ written in higarana and latin script.

```
:nihongo lemon:canonicalForm [
  lemon:writtenRep "日本語"@ja-Jpan ;
    isocat:transliteration "にほんご"@ja-Hira ;
    isocat:transliteration "nihongo"@ja-Latn ] .
isocat:transliteration rdfs:subPropertyOf lemon:representation .
```

Figure 1: Example 15 from the Lemon Cookbook

Because the literals cannot be the source of a link, the node ⟨日本語⟩ cannot be linked to entities representing each of its three constituent Chinese characters. For that, another node linking multiple resources and literal nodes is required. This makes creating and expanding the graph difficult: each modification requires the creation of multiples intermediate nodes and properties. This also changes radically the structure of the RDF molecule (Ding *et al.*, 2005) under modification. Additionally, this generates many layers of indirections that have to be handled by client applications consuming the graph. The creation of n-ary relationships also requires additional nodes and properties (W3C, 2006) because every relation in RDF is binary.

RDF does not have annotations (Lopes *et al.*, 2010), which is problematic to encode some common cases encountered in lexicography (*cf. infra*). Graph databases such as Neo4J (Robinson *et al.*, 2013) provide a similar mechanism. Finally, reification — an edge used as source or target of another link — is notoriously difficult and controversial (Trame *et al.*, 2013).

3. The Graph Model

The data model we propose is based on the notion of hypergraph (Lecailliez, 2016). A similar system is described by Williams (2000) under the name associative model of data (AMD). Despite their resemblances, this model is not directly built on AMD. Identical core concepts in both models are the use of only two kinds of entities (item/node and links) and the ability of a relation to link other relations in addition to nodes. There are however key differences in the model presented here such as the arity of links (always three in the AMD but unconstrained here) and the absence of relationship between types.

Type System

The graph is made of two kinds of objects: vertices and edges. Each object instance is associated with a type, which is an aggregate of multiple key-value pairs. There are no

predefined types of vertex or edge. The six defined properties are: Name, Identifier, Description, Object Kind, is_oriented and is_direct_content.

The **name** of a type is destined to human lexicographers and developers; it should be short yet informative. This property is never used by software to test the equality of two types. The **identifier** property is used to that end. A type identifier could be shared between projects. The **description** is where potentially long explanations are made to describe how a node content must be written and what semantic an edge carries. It may address concerns of lexicographers such as the transcription system to use, as well as developer problematics like the encoding of multimedia content. The **object kind** is either edge or vertex.

The last two properties have only meaning depending on whether the type describes a node or an edge. A vertex type with a **is_direct_content** set to true indicate the node's value can be displayed to the end user of a dictionary application without any further processing.

Node Instances & Atomicity

A node instance aggregates the following values: a type, a language tag and a list of indexed strings. In addition, the internal value of a node is stored in the *Content* property; it cannot be referenced directly from another node of the graph, nor can it contain a reference to a graph object. Nodes are hence atomic values in respect to the graph. This is an important property because (1) it allows a node to be safely removed from the graph without leaving dandling pointers and (2) any link between data must be explicitly declared with a relationship instance; this enables its automatic discovery and processing. This is contrary to the (Polguère, 2014b) model which explicitly makes use of non-atomic lexical nodes.

Atomicity allows complex content to be stored in vertices. For example, information about vowel devocalization in a Japanese word or multimedia content such as an image can be added this way. Most of the time however, simple lexicographic data can be stored as a string which can be displayed directly to the dictionary user.

Relationships

A link indicates that two or more graph objects are linked by a given relationship denoted by its type. The model currently defines three kinds of edges: non-oriented edge, oriented edge and hyper-edge. The difference between oriented and non-oriented links is semantic: an oriented instance of a relationship means it has meaning in only one direction. If the relationship can be interpreted backwards, the *Description* field should explain the associated semantic. In the prototype, a single class implement the three kinds of edges.

Hyper-edge — a link between more than two objects — is mainly motivated to enable representation of information of dictionaries where Chinese characters are present. In these languages the minimal bilingual entry is made of a word of the source language written in two forms and a word or expression of the target language. One of the graphical source forms contains Chinese characters while the other is written in a script not ambiguous about its pronunciation. For example, a minimal triplet entry for the word "birthday" in Japanese is (誕生日, たんじょうび, birthday) or (誕生日, tanjōbi, birthday) if we use the Latin script to transliterate the word.

Annotations

Not every situation benefits from being modelled as a node or an edge. Typically, information such as part of speech would have a huge number of linked nodes in a real dictionary. Because such nodes change the topology of the graph, they don't play nicely with an automatic display system such as the one presented in the next section. Annotations solve this issue by providing an alternative way to associate data with graph objects.

An annotation is a triple of strings (namespace, key, value) that can be associated to any object of the graph. It offers great flexibility of modeling but less automatic processing capabilities. The namespace allows multiple keys with the same name, for example two *freq* properties indicating the frequency of a word that were computed using different corpora.

4. Application Architecture

Using a graph doesn't provide much benefit by itself if not exploited by software. In fact, it creates problematics such as how to display the graph to the end user. We propose in this section an application architecture that uses a graph and displays it in a traditional fashion.

The Figure 2.a illustrates the architecture of a mobile application. The application package is made of three parts, one of them being mandatory. The required part (2) is composed of (a) the graph model implementation, (b) the page composer which dynamically generates dictionary pages and displays components for (c) vertexes and (d) relationships. A file (e) or code section declares the associations between graph types and display components.

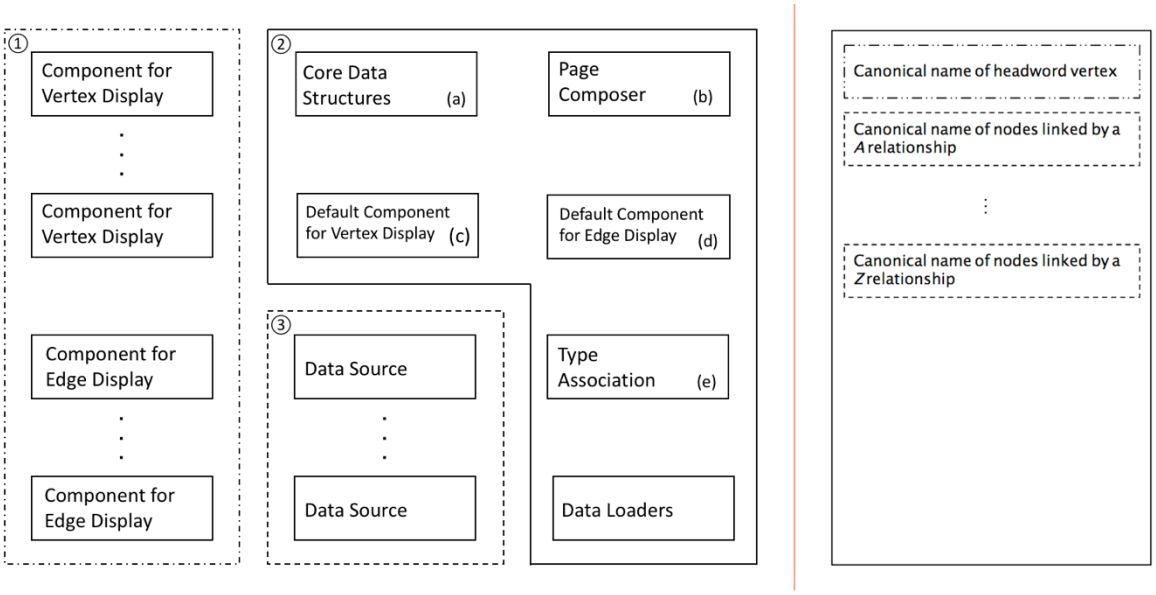


Figure 2: a. Application Architecture; b. Page Generated by the Page Composer

Types for which no custom display component (1) association is declared are handled by the default (c) (d) components. This method allows modular development and incremental additions to the dictionary. User interface is not declared in a single hard-to-maintain file. Finally, data files are bundled (3) with the application or can be retrieved from the network.

Page Generation

Dictionary pages are generated on the fly by the page composer for a given "headword vertex". The head vertex content is displayed at the top of the page (Figure 2.b) by the default vertex presenter or a custom interface component. All neighboring nodes are grouped by relation type and each group is displayed by an instance of the default edge presenter or a custom component.

The behavior is the same with any number of nodes and relationship types. New types can be added to the graph at runtime, it does not affect nor break the page composer. New data is handled by the default display component if their content can be displayed directly.

These are the two key points that allow the statement (1) made in section 2 concerning the immutable business logic of the application.

In addition to displaying vertex content, the default relationship presenter implements a navigation behavior: a click listener is added to each text data displayed, which changes the current headword with the target node associated with the data being clicked. By changing the current vertex head, a new page is generated and displayed. The user can navigate through the whole graph using this mechanism.

Custom Display Component

Finally, the composer can make use of self-contained display components that interpret a node content to produce a rich formatting. The formatting can include colors, multimedia elements and custom click behavior. Figure 3 shows two pages that use different components for data display instead of the generic one.

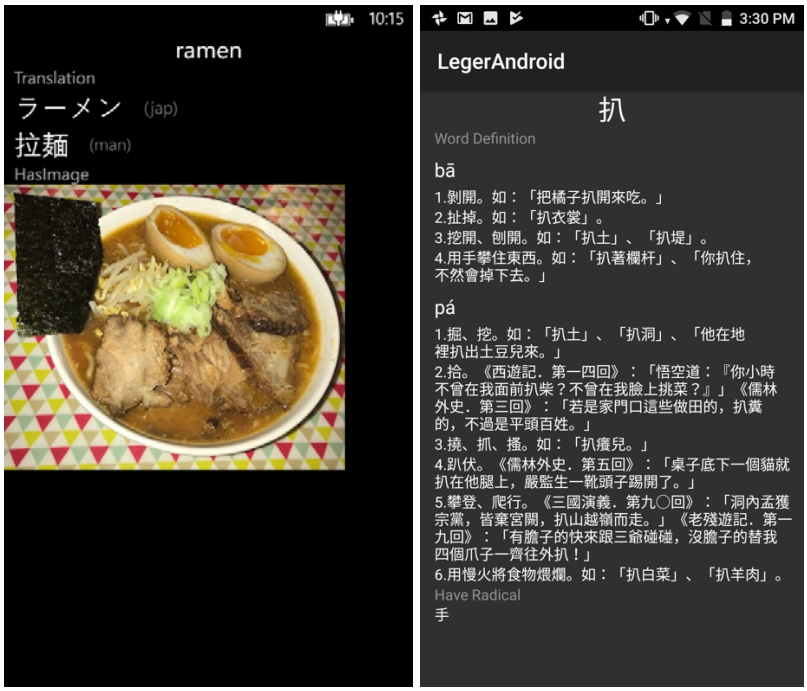


Figure 3: Use of Custom Components in (a) a Windows Phone and (b) an Android application

In Figure 3.a, the first component displays the content of nodes reached by the *Translation* relationship. It appends the language of the node in gray and between parentheses after the data itself. Another one is used to display an image related to the node reached by a *HasImage* link. In Figure 3.b data is taken from the (MoE, 2015) dictionary. The *Word Definition* relationship uses a custom UI component while the *Have Radical* does not.

Default and custom display components are the way the statement (3) of section 2 is held true. The final user interface is broken down in various and totally independent software classes. Additional components can be added without any need for changes to existing ones when a new data type is added to the dictionary. If this type is just a string that can be directly displayed, the creation of a component is no longer needed.

Prototype & Data

Distinct implementations of the application architecture were made for Android and Windows Phone. Each of them relies on a C# library that implements the graph model presented in section 3. The library is also used in other lexicography related programs.

We tested the model with various data sets, most notably data extracted from the Revised Mandarin Chinese Dictionary, an unofficial Japanese vocabulary list for the highest Japanese Language Proficiency Test level, and the Jibiki Project (2015), (Mangeot, 2016). A graph of Chinese and Nôm character decomposition was built with data from the Kanji Database Project (2015).

Known Display Issues

The graph connectivity can have a bad effect on display. Some nodes are connected to a huge number of vertices, so their display can be overwhelming for the user. There is no obvious and automated fix for this case: the problem must be tackled during the creation of the dictionary. Node pruning can be achieved by different algorithms that take various information in account, but it should be done statically if it uses intensive computations. The best solution in term of quality of entries is to manually annotate the most interesting entries, display a limited number of linked vertices and provide a button to load more at the user's will.

5. Conclusion and Future Research

The present paper has introduced a lightweight hypergraph model that alleviates the difficulties that existing frameworks based on RDF impose for modeling graphs. Three issues are specifically addressed: first, the need for an annotation system. Secondly, the model allows relation of n-arity, that is particularly useful when dealing with Chinese characters. Thirdly, it provides a built-in reification mechanism.

Moreover, it lays the way to create dictionary applications that can be updated easily to support additional languages. In particular, the prototype we built on the described abstract application architecture is robust to changes in the dictionary data: new types — which encode part of the dictionary micro-structure — can be added or deleted without breaking the application. It is capable of discovering and displaying string data without further modification. Elaborate display of data can be added to the software with new components; no change is required to any existing user interface files.

Finally, an edge that may recursively link nodes or relationships allows expressing complex lexicography situations. It could allow for example the user to navigate to the meaning of a word it has in the context from a higher level lexicographic construct such as a Chinese proverb. Future work will be done to provide a demonstration of this capability.

One issue though is the lack of a constraint system for relationships which could hinder the discovering capabilities of a client processing the graph (for example a program building a SQL database from a graph file). This point may be addressed in a future revision of the model.

6. Acknowledgement

This paper was partially written with the support of the MSMT10925/2017-64-002 scholarship grant from the Czech Ministry of Education, Youth and Sports.

7. References

Declerck, T., Wand-Vogt, E. et Mörth, K. (2015). Towards a pan european lexicography by means of linked (open) data. In Kosem, I., Jakubícek, M., Kallas, J. et Krek, S., editors: *Proceedings of eLex 2015. Biennial Conference on Electronic Lexicography (eLex-2015), electronic lexicography in the 21st century: Linking lexical data in the digital age*. Trojina, Institute for Applied Slovene Studies/ Lexical Computing Ltd., Trojina, Institute for Applied Slovene Studies, Ljubljana.

Ding, L., Finin, T., Joshi, A., Peng, Y., Da Silva, P. P., McGuinness, D. L. (2005). Tracking RDF Graph Provenance using RDF Molecules (revision 2). Technical report *TR-CS-05-06*, April 2005. Accessible at: https://ebiquity.umbc.edu/_file_directory_/papers/178.pdf

Hashimoto, J. M. (1973). Current Developments in Sino-Vietnamese Studies. *Journal of Chinese Linguistics*, 1-26. Accessible at: http://www.jstor.org/stabe/23752818

Jibiki Project. (2015). Accessed at: http://jibiki.fr. (20 May 2017).

Kanji Datase Project. (2015). 字形ＩＤＳデータ. [jikei IDS dēta] Accessed at: http://kanji-database.sourceforge.net/ids/ids.html (20 May 2017).

Lopes, N., Zimmermann, A., Hogan, A., Lukácsy, G., Polleres, A., Straccia, U., & Decker, S. (2010, June). RDF needs annotations. In *W3C Workshop on RDF Next Steps*, Stanford, Palo Alto, CA, USA.

Lecailliez, L. (2016). Pour une modélisation de dictionnaires de japonais sous forme de graphe. [Towards Graph Modeling of Japanese dictionaries] Master thesis, Paris Diderot. Accessible at: https://louis.lecailliez.net/dl/memoire_m2_jap_Lecailliez.pdf.

Lecailliez, L. (2017). Preliminary Thoughts on Issues of Modeling Japanese Dictionaries Using the OntoLex Model. In Horák, A.; Rychlý, P.; and Rambousek, A., editor(s), *Proceedings of the Eleventh Workshop on Recent Advances in Slavonic Natural Languages Processing*, RASLAN 2017, pages 11-19, 2017. Tribun EU. Accessible at: https://nlp.fi.muni.cz/raslan/raslan17.pdf.

Mangeot, M. (2016) Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary. *International Journal of Lexicography 2016*; doi:10.1093/ijl/ecw035; 35 p.

Mangeot, M., Sérasset, G. (2001). Papillon Lexical Database Project: Monolingual Dictionaries and Interlingual Links. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*. November 27-30, 2001, pages 119–125, Tokyo, France.

McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Pérez, A. G., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2010). The lemon cookbook. Technical report, Monnet Project (June 2012), www.lemon-model.net.

McCrae, J., Spohr, D., Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In: Antoniou G. et al. (eds) *The Semantic Web: Research and Applications*. ESWC 2011. Lecture Notes in Computer Science, vol 6643. Springer, Berlin, Heidelberg.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41.

MoE. (2015). Revised Mandarin Chinese Dictionary of the Ministry of Education of Taiwan. [教育部重編國語辭典修訂本]
Accessed at: http://resources.publicense. moe.edu.tw/dict_reviseddict_download.html

Polguère, A. (2012). Lexicographie des dictionnaires virtuels. In Apresjan, Y., Boguslavsky, I., L'Homme, M.-C., Iomdin, L., Milićević, J., Polguère, A. et Wanner, L., eds:

Meanings, Texts, and Other Exciting Things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovič Mel'čuk.

Polguère, A. (2014a). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418.

Polguère, A. (2014b). Principes de modélisation systémique des réseaux lexicaux. Principes de modélisation systémique des réseaux lexicaux. In *TALN* 2014 (pp. 79-90), Marseille.

Powers, S. (2003). *Practical RDF*. Sebastopol: O'Reilly & Associates.

Robinson, I., Webber, J., Eifrem, E. (2013). *Graph databases*. O'Reilly Media, Inc.

Trame, J., Keßler, C., Kuhn, W. (2013). Linked data and time–modeling researcher life lines by events. In *International Conference on Spatial Information Theory* (pp. 205-223). Springer, Cham.

W3C. (2006) Accessed at: https://www.w3.org/TR/swbp-n-aryRelations/. (20 May 2017)

Williams, S. (2000). *The associative model of data*. Second Edition. Lazy Software.

Entryword Choice in Bilingual Dictionaries in the Digital World: New Challenges

Mats-Peter Sundström
Swedish translation unit, European Parliament
matspeter.sundstrom@europarl.europa.eu

1. Expanding an existing dictionary with new lexemes

Lexicography in the Digital World comes with unprecedented possibilities, such as enabling the lexicographer to expand a dictionary well-nigh indefinitely. This is a subject lending itself to all but endless elaboration, but this paper will confine itself to discussing the need to add further headwords to a major (75 000 + entrywords) bilingual dictionary having English for its source language (and basically any target language), as a result of recent developments particularly in media language. Further, it should be mentioned that the bilingual dictionaries that form the subject matter of this paper are supposed to be electronic ones, although they may also appear in printed version. Reference will here be made to Lew & de Schryver 2014; "The most popular and broadly used term for digital-media dictionaries has been electronic dictionaries, sometimes abbreviated as e-dictionaries" (Lew & de Schryver 2014, 342).

At this initial stage, mention must be made of another very pertinent observation made by Lew & de Schryver regarding the altogether different operational conditions for lexicographers involved with electronic, as opposed to printed dictionaries: "In the print-dictionary age, one of the main motivations for updating dictionaries was to accommodate new vocabulary [...]. This usually involved painful decisions as to how to accomplish this without the printed volumes overshooting their target size [and] in the end the editors usually had to grapple with the dilemma of what to sacrifice in order to make space for new items. The digital revolution has changed that, and new items are in fact very rarely removed when digital dictionaries are updated." (Lew & de Schryver 2014, 345).

The issue of neologisms no doubt makes itself felt as a major consideration of lexicography. Relatively speaking, however, the number of perfectly new lexical items ("full neologisms") is restricted. From the field of gender equality, mention may be made, for instance, of *mansplaining* and *me too-generation*. Perhaps of greater interest is what we for the purposes of this paper may call "partial neologisms". Such items of language basically involve a change of register, notably from LSP to general language. Actually, for the purposes of the present paper, the notion of neologism ought to be broadened to include also semantic shifts and shifts of register.

2 Neologisms where LSP elements are taken over into general language

Among elements deserving enhanced attention in major bilingual dictionaries involving English as their source language may be noticed elements of LSP that, through media language, have come to assume the character of general language lexemes. This process is summed up by Svensén through the term of "avterminologisering", i.e. determinologisation (Svensen 1987, 299). They come from a variety of subject fields and sometimes their origin can be attributed to individual persons. Such is the case for *disruptive technology*, defined as technology that "displaces an established technology and shakes up the industry or a ground-breaking product that creates a completely new industry" (Rouse 2016) and originally coined by professor Clayton M. Christensen. Several others have been derived from the terminology of their respective technical fields and subsequently found their way into media language. Think of *coping strategy* (derived from psychology; rather self-explanatory), *decarbonisation*

(from environmental affairs, "the reduction or removal of carbon dioxide from energy sources"), *rebound effect* (from economics, the cancelling out of a desired effect by subsequent undesired developments) or *troll* (initially a creature of Norse folklore, later a malevolent internet user).
The subject of semantic shift in established lexemes offers some fascinating glimpses as to why bilingual dictionaries involving English as their source language may be in acute need of complementing, owing to recent developments notably in media language. To delimit this potentially unlimited subject, examples will be drawn from two thematic fields: 1) equal rights and 2) environmental affairs.

2.1 Examples from the field of equal rights

In the first field, the development undergone by the noun *gender* is rather obvious. From having simply designated sex as a social, as opposed to a biological construct, it has now more or less superseded the word *sex* as the general designator of maleness or femaleness. Consider for instance English- language questionnaires, where the caption above the question bearing (usually) the alternatives of male and female is *gender* rather than *sex*, the latter these days being reserved for acts of sexual behaviour. It may be wondered whether all bilingual dictionaries involving English for a source language have taken account of this development – or indeed and perhaps even more, whether bilingual dictionaries featuring English as their target language have done so. Likewise, owing to language development, the adjective *gay* can hardly anymore be used in its primary sense of "cheerful", "merry", nor can *intercourse* readily be used in reference to "social interaction" or "social contacts", at least not unless preceded by the qualifying word *social*, as in "social intercourse". – Appropriate warnings to non-English dictionary users will become necessary. For after all, the line in the well-known American song *My Old Kentucky Home* where one line goes *The young folks roll on the little cabin floor / all merry and happy and gay* will most likely, to present-day readers, give rise to interpretations very foreign to the author's original intentions ...

Then, there are more subtle shifts within this area also. The noun *rainbow*, notably when used adjectivally, provides telling examples, referring to the rights of sexual minority communities, such as a *rainbow family*. As was pointed out by Swanson: "Gilbert Baker, an artist [...] first created the Rainbow Flag in 1978" to symbolize the gay rights movement (Swanson 2015) and subsequently this new use of the word rapidly spread.
Against this background, the description of the Republic of South Africa as the *rainbow nation*, originally initiated by the Most Reverend, Archbishop Desmond Tutu immediately after the fall of apartheid may come across as potentially misleading. Also, to modern English native users, the noun *chauvinist* probably coincides primarily, not with extreme patriotism, but with someone who is gender insensitive. Finally, *pride* in current media language frequently connects with the rights of homosexuals, notably in compound words such as *pride marches/events/festivals*.

Gender equality issues have given rise to a welter of new expressions, frequently found for instance in texts from international bodies such as the United Nations and the European Union. As an example may be taken *gender budgeting* i.e. paying attention to issues relating to equality between men and women in connection with budget planning, or *gender mainstreaming*. i.e. integrating gender equality aspects in all policies and other work, or why not the adjectives *gender sensitive* and *gender responsive*, both of which are semantically neatly summoned up in L. Aquilar's definition "Being gender-responsive means going beyond acknowledging gender gaps and really doing something about the discrepancies" (Aquilar 2015). Interestingly enough, as a qualifying first element in compound words, *gender* has often come to acquire the sense of *gender equality*, as is attested by examples such as *gender considerations*, *gender perspective* or *gender policies*.

2.2 Examples from the field of environmental affairs

Over now to the second thematic field, that of environmental affairs. The first example that springs to mind is obviously that of *sustainable*, as primarily in *sustainable development*, a household word of development policy since the Brundtland report of 1987, defined as a development that "meets the needs of the present without compromising the ability of future generations to meet their own needs" (UN reports 1987). Most probably established as a term in most major languages by now, this expression, interestingly enough, sometimes has semantically somewhat different equivalents in different languages, such as *kestävä kehitys* in Finnish (literally "lasting" or "durable"), whereas in Polish the corresponding concept is *zrównoważony rozwój* (literally "balanced development"). Since then the adjective *sustainable* has undergone a marked shift of meaning, referring as it does, very often to activities and processes that are environmentally sound, i e, do not cause irreparable harm to the environment. *Sustainable construction*, for instance, refers not so much to the physical strength and stress tolerance of a building, but rather to its being as little damaging as possible to Earth's climate and resources, both during construction and operation.

Something similar is true of the adjectives *green* (in the sense of *environmentally friendly*) and *organic* (in the sense, not as the opposite of *inorganic*, but rather, produced in a manner respectful of the environment and its requirements). In connection with the adjective *green* mention may also be made of the colour name *blue*, referring to sea-related matters in compound words such as *blue growth* (defined by the European Commission as *the long term strategy to support sustainable growth in the marine and maritime sectors as a whole*) (European Commission). The ongoing debate about climate changes has enlarged this list of semantically shifted lexemes with items such as *fossil free* (meaning not "free from fossils" but "not using fossil fuels"), *low carbon* (not referring primarily to something having a low carbon content, but rather to its causing feeble emissions of carbon dioxide), and *renewable* (used as a noun meaning not just anything capable of being renewed, but rather a source of renewable energy).

Needless to say, neologisms and shifts of meaning are not confined only to subject fields such as equal rights or the environment. A graphic illustration is provided by the word pairs *transparent/transparency* and its antonym *opaque/opacity*. The former pair, from having meant simply "through-lookable" has to an increasing degree come to be used in the sense of "frank, aboveboard, free from (undue) secrecy", as in *transparent elections*, *transparent procedures* and so on, whereas the latter pair denotes the opposite. Worth noticing in this connection is the international organisation Transparency International and the declaration on its home page: "We want to make decision-making in the EU as *transparent* as possible" (Transparency 2018).

Evidently, neologisms and semantic shifts of the kind referred to above occur, not only in major languages such as English but also in minor languages, such as my native language of Swedish. Sure enough, as the Finnish aphorist Paavo Haavikko once put it: writing in a minor language does not prevent you from saying things of major importance. Since we have been discussing environmental affairs in the previous passage, the introduction of a Swedish word from this field appears relevant, i.e. the noun *miljötänk*. Formed as a portmanteau word from the noun *miljö* ("environment") and the verb *tänka* ("to think") it basically means "considering/caring for the environment"). This word is often untranslatable into English, as the English direct equivalent *environmental thinking* would function as a substitute only on a limited number of occasion. Rather, a Swedish phrase such as *vi hade inte miljötänk på den tiden*, literally "we had no environmental thinking back in those days" would idiomatically translate into English as *the environment was not a matter of great concern to us back in those days*, to mention just one translation option among many others.

One of the main problems of rendering English lexemes of the kind shown into other languages may derive from the impossibility of finding or inventing one-word equivalents, necessitating cumbersome, circumlocutory expressions that are descriptions rather than equivalents. A telling illustration may be provided by the term *compliance officer* which, according to the European Union database IATE officially translates into Finnish as *säännösten noudattamista valvova virkamies* i.e. *rules observance supervising official*. In such cases, very often there is not just one established translation option, but several, and in the relative freedom from constraints of space provided by the digital age enables dictionary editors to vastly enhance the presentation of these options within the confines of a dictionary article.

3. Why include examples of "organisolects" in bilingual dictionaries?

Above was presented a perhaps daunting multitude of examples culled most notably from texts related to the work of the European Union (after all, the present author has been a language officer for the European Parliament for more than twenty years), along with other international organisations. Now it is high time to find a synthesis, to explain why in the first place such considerations matter from lexicography's point of view. The reasons are fourfold:

1. Our world is increasingly more dominated by multilateralism, manifested in the work of numerous international fora and organisations, whose predominant working language is English. By this is meant organisations such as the United Nations and its special agencies, the European Union, the IUCN (International Union for Conservation of Nature), and non-government organisationes such as Greenpeace, Human Rights Watch, Transparency International, to mention just a very few.

2. Their work influences the English language, giving rise to neologisms and shifts of meaning in existing lexemes, a process obviously reflected in English-language mass media and thus, spreading out from a rather narrow circle to reach the general public, both native speakers of English and persons to whom English is a foreign language.

3. Above all, in order to facilitate the latters' comprehension of modern English media language, lexemes of the kind presented above merit inclusion in bilingual dictionaries having English for their source language. This is all the more important in situations where established equivalents (whether one-word or multiple-word equivalents) have evolved in the target languages. As far as these international language "buzzwords" are concerned, such must be the case notably for the Chinese language, given its status as one of the six official languages of the United Nations.

4. Today's ubiquitous digitalisation has been an immense enabling factor for lexicographers' excerpting work. Search engines lead us, not only to individual words but also to the contexts in which they occur, and sometimes to veritable dictionary-style explanations (such as can be found upon giving a search command *What is the meaning of X*).

All this said there are compelling reasons why lexicographers should mobilise their resources to make sure that dictionaries reflect not only "traditional words" and "traditional senses" but also emerging ones. In this connection, a counter-argument may be advanced, i.e. the uncertainty as to whether these linguistic innovations will become lasting elements of language. This argument, however, is somewhat specious. With modern technology, lexemes may easily be either deleted or supplied with a label indicating their obsolescence. Evidently, many lexemes of the kind described are characterised by a marked degree of epochality, i.e. being closely connected with a given epoch in the past and now fallen into disuse. Consider *detente or women's lib* or the once much-loathed *finlandisation* all strongly redolent of the 1970's. Admittedly, they serve few if any functions today, but they act as a testimony of language in days not so long gone by, and thus may well deserve being preserved in dictionaries, if provided with labels such as "historical" or similar ones. And, as was pointed

out right at the beginning of the paper: expanding an electronic dictionary these days seldom presupposes deleting obsolete items.

In this context a word of caution imposes itself. In the words of Svensen, there are ghost words, i.e. words that really do not exist, save in dictionaries (cf Svensén 1987, 29). Evidently they may consist either of single words or multiple phrases, such as (alleged) idioms. Of the latter, Moon points out that "items like *kick the bucket* and *rain cats and dogs* may be talked about more than actually used in current English, so their omission could be seen as entirely healthy" (Moon 2016, 322). Another example can be found in a source rather more uncommon to the present audience, i.e. a major dictionary between Finnish and Swedish where an allegedly Swedish collocate phrase is given *hon är ful men kul* (Cannelin 1986), in English "she is ugly but nice". All researches, however, indicate that this phrase is a nonce expression, thought up by the dictionary editor himself. This serves to corroborate the words of Hanks, to the effect that "pre-corpus dictionaries are full of unnatural invented examples" (Hanks 2012, 232).

4. Expansions with elements from informal language

Then, on a very different level comes the ever-increasing prevalence informal language items in English-language media. For one thing, media prose in English very often lacks the aversion to slang as can be found in several other language areas. Consequently, in English news reporting it is not uncommon to find expressions such as *gung-ho on* (instead of, e.g. *keen on*), *oomph* (instead of e.g. *momentum*, as in *the economy ran out of oomph*), *pizzazz* (instead of, e.g. *impetus*, as in *adding pizzazz to the economy*).

At this point may be warranted a remark in passing as to why these decidedly informal lexemes should even find their way into journalist prose. The reason, however far-fetched it may seem, may be found in a perfectly different view of a journalist's task and vocation in the Anglo-American cultural sphere as opposed to many other cultures. An Anglo-American journalist is seen, first and foremost, as a kind of society's watchdog, always prepared to alert the public to abuses of power, underhanded dealings and so forth. This position then also implies permanently assuming a kind of cavalier, irreverent attitude to society and its makers and shakers, also reflected in purely language terms. Additionally comes the constant desire to keep the reader from becoming bored, and to that end, constantly attempting to amuse the readership.

For instance, the message contained in a statement like "Defence high command desires more firepower in exchange for tax payers' money spent" may well be expressed with a phrase like *Army top brass want more bang-bang for every buck taxpayers cough up.* In many other cultures, such a turn of phrase would be considered both frivolous and flippant, not having any place in serious news reporting. In the place where it originated, however, it will most likely be perceived as an example of brisk and brazen, that is commendable, journalism.

Another reason behind the increase in informal English lies in the simple fact that the uptake of social media markedly increases the number of people expressing themselves in writing. And since so much of social media exchanges take place in English, there is bound to be ample room for slang, jargon and the like. Additionally, it is worth pondering whether or not the commonest acronyms of sms-ing and social media should merit inclusion in the lemma list, such as *asap*, meaning "as soon as possible", *lol* = "lots of laughter" or *omg* (despite its potentially offensive character to believers) = *oh, my God*, the latter even found in compound words such as *omg-worthy*.

5. Pragmatics: usage operating on target language, not source language terms

This said, we arrive at the somewhat dodgy issue of pragmatics. In a way, it is up to question whether this issue should be touched upon in a dictionary at all. For surely, a dictionary, whether monolingual or bilingual, is about meaning, which gives rise to the question as to how this rather obvious fact can be reconciled with Čermák's statement to the effect that "pragmatics is best handled outside the description of meaning, if possible" (Čermák 2015, 354). However, as was pointed out to the present writer by professor Shigeru Yamada: "electronic dictionaries are virtually unlimited in terms of the information they can provide", so why should we then omit to take advantage of these possibilities? (Yamada 2015). As the present writer pointed out at the Ninth Asialex Conference in Hong Kong 2015, dictionary articles can, where appropriate, "be supplied with pragmatics boxes" (Sundström 2015).

In brief then, such a structure could take the form of a framed-in box, located perhaps after the dictionary article, supplying additional information about the entryword. For instance, a foreign English learner with either French or Polish for first language may find it useful to know that although it is perfectly natural to say *selon moi* in French or *wedlug mnie* in Polish, the corresponding word-for-word translation into English as *according to me* would sound non-native, or worse still: pretentious. Likewise, a person with Swedish for first language may find it relevant to learn that although the English language does have separate designations for different kinds of headdresses (hats, caps, hoods and so forth), in actual language practice almost any headdress may be referred to as a hat. Incidentally: the present author was no little surprised when he had once dropped his peaked cap on the church floor and a young English boy came up to him with the message: "Sir, please don't forget your hat!". Somewhat similar was his reaction when reading about "Santa Claus hats"... for most certainly, to him, viewing the world through the prism of the Swedish language, Santa wears a hood and nothing else. Even so, the world seen with an English-speaking person's eyes takes on a different appearance, something well worth noticing in a dictionary.

Another cue to this theme may be provided by the way the names of tree species are handled in common usage. A photo in the European newspaper Politico recently came with the caption: "Workers cutting down a large spruce *tree*" (my italics). Again, approaching the issue through the prism of the Swedish language is likely to cause confusion. Surely everyone knows that a spruce is a tree, so why such a pleonastic expression in English as a *spruce tree*? Notwithstanding this possible objection, the fact remains that English prefers expanding the construction by adding the word *tree* (surely some of us remember the hit song of the seventies where a yellow ribbon was supposed to be tied around the "ol[d] oak tree" rather than just around the "ol[d] oak". So far, the present author has seen no mention to this effect in any dictionary consulted by him.

Pragmatic comments may include information to the effect that in actual language usage a word may be used in additional senses than the "official" ones, in spite of the fact that such practices may be frowned upon by language planning bodies. The word *leverage* when used as a verb offers a graphic illustration to this tendency. As the following comment suggests "Leverage as a verb is either (a) an unnecessary neologism meaning 'to use'; (b) a term for investing with borrowed money, or any economically equivalent act" or " "*Leverage* is an unnecessary verb introduced to make statements sound more technical than they are. Some options to consider are *enhance, use, exploit, utilize and employ*"(englishstackexchange). However, without prejudice to the opinion of language planners, persons to whom English is a second language will most likely have a legitimate need to learn also about such uses which, although perceived as formally incorrect nonetheless exist in the real life of language. For after all, such language items can always be provided with a label such as *krytykowane*. [= 'criticised') as in PWN-Oxford Polish–English Dictionary (PWN-Oxford 2012).

Finally, collocations are definitely an area where pragmatics come into play and where digital technology with its exponentially expanded search possibilities will be of great avail. For instance, lexicographers may find new examples of cases where source and target language collocate differently with their respective lexemes. To illustrate: a word-for-word translation from the Swedish language would result in a phrase such as **go in advance of events* while a possible English corresponding expression, gleaned from an English news magazine could be *run ahead of events*. Given today's search engines lexicographers can rapidly either vindicate or invalidate such hypotheses.

6. Possibilities are infinite but dictionaries are not – what is the conclusion?

Decidedly, no dictionary in the world, whether bilingual or monolingual, can aspire to cover even a fraction of the ever-changing spectrum of novelties in language, nor should it try to do so. In addition, a caveat is warranted. As Lambert put it "although electronic dictionaries have essentially unlimited space, lexicographers still need to balance practical utility against the time and expense of researching and writing entries" (Lambert 2017, 39 f). Yet, modern digital technology has vastly improved our investigative possibilities, whether in the shape of corpus queries or search engine research. In the mind of the present writer, lexicographers should not fail to make use of these new options for complementing entryword bases and bridging the possible gap between what is shown in dictionaries and what occurs in actual language usage.

To mention but one possibility, potentially of particular interest to those involved with English-Chinese lexicography. Granted that Chinese is an official language of the United Nations, every single resolution from that organ has to appear in Chinese. To all appearances, this would be a veritable gold mine for finding Chinese equivalents for words and expressions used within a UN context. Similarly, since Japan since 1996 holds observer status within the Council of Europe (not to be confused with any European Union body), there is reason to believe there could be no small amount of texts relating to European affairs translated into Japanese, complete with term definitions and all, mostly from English but sometimes from French originals. Although it wholly escapes the present author's knowledge to what extent sources of the above kind have been drawn upon, the suggestion is definitely worth making.
So the words of Goethe *in der Beschränkung zeigt sich der Meister*, or, in John Irons's congenial English translation from 2011: "The master shows himself first in confinement" might thus serve as a kind of motto for lexicography in today's digital world. In sum then: if we are editing dictionaries, let us not shun away from complementing our products, mindful, however, of the lexicon's ever-changing protean character and language's intrinsic impossibility to lend itself to complete representation in a dictionary, whether in print or electronically.

Sources

Cannelin 1984: Cannelin, Knut, Hirvensalo, Lauri, Hedlund, Nils: *Finsk–svensk storordbok*. (Finnish–Swedish General Dictionary). Porvoo 1986.

Čermák 2015: Čermák, František: *Semantic Explanations and Entry Structure in Idiom Dictionaries*. International Journal of Lexicography Vol. 28 No 3, pp. 353–359.

Hanks 2012: Hanks, Patrick: *The Corpus Revolution in Lexicography*. International Journal of Lexicography Vol. 25 No 34 pp. 398–432.

Lambert 2017: Lambert, James: *Ornithomy and Lexicographical Selection*. International Journal of Lexicography Vol. 30 No 1, pp. 39–62.

Lew &de Schryver 2014: Lew, Robert; de Schryver, Gilles-Maurice: *Dictionary Users in the Digital Revolution*. International Journal of Lexicography Vol. 27 No 4 pp. 341–359.

Moon 2016: Moon, Rosamund: Idioms. A View from the Web. International Journal of Lexicography Vol. 28 No 3 pp. 318–337.

PWN-Oxford 2012: Wielki *Słownik* Polsko–Angielski. Polish–English Dictionary (editor in chief: Jadwiga Linde-Usiekniewicz). Warszawa 2012.

Sundström 2015: Sundström Mats-Peter: *Bilingual dictionary and encyclopedia – where goes the line? Cultural specificities as sources of dilemmas and solutions proposed*. Paper presented at the Eight Asialex Conference, Hong Kong 2015.

Svensen 1987: Svensen, Bo: *Handbok i praktisk lexikografi*. Stockholm 1987.

Yamada 2015: Oral communication from professor Shigeru Yamada.

europa.eu/rapid/press-release_MEMO-13-615_en.htm: *Blue Growth strategy to create growth and jobs in the marine and maritime sectors gets further backing*.

UN Reports 1987. *Report of the World Commission on Environment and Development: Our Common Future*.

https://english.stackexchange.com/questions/45234/difference-between-leverage-and-utilize *(accessed on Mar. 14 2018).*

https://transparency.eu *(accessed on May 17 2018), anonymous.*

https://whatis.techtarget.com/definition/disruptive-technology (accessed on May 17 2018), posted by Margaret Rouse, December 2016.

www.huffingtonpost.com *Stop Being So Sensitive! The Shift from Gender Sensitive to Gender-Responsive Action* (accessed on Mar. 13 2018), posted by Lorena Aquilar Oct. 8 2015.

www.washingtonpost.com *How the rainbow became the symbol of gay pride* (accessed on Mar. 13,2018), posted by Ana Swanson June 29 2015.

Learner's Pharmaceutical Dictionary: the Question of Content and Design

Matyushin A.A. & Markovina I.Yu.
Sechenov First Moscow State Medical University. Moscow, Russia
matyushin@1msmu.ru

Abstract

Different kinds of lexicographical sources are used by learners throughout their university years. Among them bilingual LSP (Language for Specific Purposes) dictionaries play a unique role as they serve both as a source of professional knowledge (at least to a certain extent) and an educational means to provide learners with a profound understanding of the foreign language of their field. According to the Function Theory of Lexicography a concept of a specialized learners' dictionary relies not only on the needs of specific users but also on the specific social situation. To our knowledge, no such situations have been identified and described for the pharmacists so far, hence the lack of a specialized learner's dictionary for pharmacy students. The aim of our research was to identify thematic domains related to the pharmacist profession and relevant professional social situations, and to describe the principles of developing a bilingual LSP learner's dictionary for pharmacy students. The most common cognitive, communicative and operational extra-lexicographic user situations were described using the methodology proposed by S. Tarp. The paper also provides a brief criticism of some of the existing pharmaceutical dictionaries and gives the authors' insights into the pharmaceutical learner's dictionary concept. We focus on the motivation for preparing such dictionary, the use of corpus-based data to establish its possible content, as well as on our vision of the optimal macro- and microstructure and user interface, with particular focus on the possibility of individualization, as only individual approach to learner's needs can bring us closer to an "ideal dictionary".

Keywords: learner's dictionary, pharmaceutical dictionary, LSP, building vocabulary

Introduction

The article presents the detailed concept for both digital and printed LSP (Language for Specific Purposes) pharmaceutical dictionary designed as a tool that can be used in a variety of situations by pre-graduate, graduate and postgraduate pharmacy students to master their knowledge of L2 as well as by pharmacy professionals and translators. The dictionary will be produced by experienced pharmacists with a second degree in translation in collaboration with professional linguists. We will briefly discuss our motivation for designing and producing such dictionary, requirements for corpus of texts that will be used to develop the dictionary content, and the dictionary microstructure that may provide optimal user experience. The function theory of lexicography and the experience of producing several bilingual medical dictionaries serve as the basis of our dictionary concept.

Background of the study

In 2016, the pharmaceutical faculty of Sechenov University underwent a much needed major reorganization. This was accompanied by tremendous changes in undergraduate curriculum: the duration of training has remained the same (5 years + optional residency), but many new courses and elective subjects were added and some of the out-of-date ones were partially reduced or completely removed. The rationale for these changes was based on employers' demand to hire graduates with a set of skills more suitable for the modern pharmaceutical industry, who would not require additional on-site training. Curriculum changes were based on a 3+ edition of Federal State Educational Standard (FSES) for the "Pharmacy" specialty, which states that after graduation any pharmacist must be able to participate in the oral and written communication both in native and foreign language in order to fulfill professional responsibilities. Therefore, the Chair of Foreign Languages (presently – Institute of Linguistics and Intercultural Communication) of Sechenov University was required to develop and implement new programs for pharmacy students.

The analysis of current needs of pharmacy students carried out by a joint group of experts, allowed to identify a gap in the teaching of professional pharmaceutical terminology. Currently, during their pre-university years students use a variety of dictionaries and other lexicographical publications (encyclopedias, reference books, etc.) in order to build up their general native and foreign vocabulary; after graduation they are exposed to a vast world of professional terminology developed, used, and maintained by regulatory authorities, pharmacopoeia committees, drug manufacturers, and researchers all over the world. However, even the reorganized curriculum does not include separate course of professional terminology. At best, students are introduced to the basic Latin terms used in medical prescriptions and a heavily fragmented terminology of a specific subject (chemistry, biology, physics, etc.) taught in the native language (Russian) whereas most of the pharmaceutical terminology nowadays is English.

This (hopefully, temporary) lack of structured terminological training can be made up for in several ways. First of all, several new textbooks and workbooks were proposed with the aim to simultaneously provide necessary information about pharmaceutical terms as well as English grammar and syntax. Development of content of these books will be accomplished in collaboration with faculty members who specialize in different areas of pharmacy; their input is required to assure that the materials are correct and suited for pharmacy students.

An additional tool for filling in terminology knowledge gap is the bilingual learner's dictionary. Bilingual LSP learner's dictionaries may provide unique opportunity to enhance both professional knowledge (at least to a certain extent) and profound understanding of the foreign language of the specific field. The lack of such dictionary for pharmacy students served the key stimulus for the development of our research objectives.

Apart from the above, our motivation to develop the concept of a learner's pharmaceutical dictionary was based on the following facts:

- In modern world there is an ever-growing demand for high quality medicines;
- Current revenue of the worldwide pharmaceutical market is about 1200 bn USD which makes it one of the most important business areas;
- English language is one of the most spoken languages in the world, is widely used as an official language, and all new pharmaceutical information is released primarily in English;
- Besides locally developed glossaries, large learner's pharmaceutical dictionaries are virtually non-existent.

The experience obtained from several projects of bilingual medical dictionaries (e.g. "English-Russian and Russian-English Medical Dictionary" by I.Yu. Markovina) and the background of teaching English language to medical and pharmaceutical students makes the task a bit less challenging. However, the concept of this new dictionary has to be adapted to new technologies, i.e. many of the dictionary design stages should be reconsidered in order to provide better user experience, both in terms of usability and suitability.

Why can't we use an existing pharmaceutical dictionary?

To make a long story short, we simply don't have any suitable learner's pharmaceutical dictionaries. Most of the existing domestically-developed dictionaries and glossaries (i.e. produced by university language departments) designed to enhance learner's skills in written communication have quite low lexicographical quality. Several large non-learner's pharmacy-related lexicographic works (ISPE Glossary of Pharmaceutical and Biotechnology Terminology, Academic Dictionary of Pharmacy, Dictionary of Pharmacoepidemiology, Dictionary of Pharmaceutical Medicine, and other) are well-constructed and their content is somewhat up-to-date; however, they are not designed to be used as learner's dictionaries as they provide only definitions for lemmas and references to other entries, but no other information usually required by the learner (e.g. translation in case of a bilingual LSP dictionary). In other words, these specialized dictionaries designed to provide knowledge about field of pharmacy do not cover users' learning needs. Besides, they suffer from serious disadvantages that diminish their role in educational process. They are designed either to be used by pharmacy professionals who require a reference book in order to aid their text comprehension and writing or by pharmacy students whose L1 is English, and the needs of students who have English as L2 are not met.

Elsevier's Dictionary of Vitamins and Pharmacochemistry stands apart from other pharmaceutical non-learner's dictionaries. Despite having entries in English, French, German, and Portuguese, and the indices in all of these languages., it provides quite odd and questionable selection of lemmas. We find overall design of the dictionary very convenient: each entry has identity number, lemma, equivalents of the lemma in other languages as well as definition and grammatical information. However, the idea of using a part of the sentence or a whole sentence (sometimes up to 10 words!) as a lemma seems doubtful. A separate field which could illustrate a specific sense or construction would be better in terms of helping the the dictionary users to understand lemmas in their texts and use them correctly.

Objectives

Basically, a developer of any dictionary, be it learner's dictionary or any other type of lexicographical works, should answer two fundamental questions (Tarp, 2008a): what should be included in the dictionary and how this content should be organized. Aiming at answering these two questions, the objectives of our study were as follows:

- to identify the learner's needs;
- to identify thematic domains related to the pharmacist profession;
- to identify relevant social situations which require the use of the dictionary;
- to propose an optimal microstructure of the dictionary;
- to describe the principles of compiling a bilingual LSP learner's dictionary for pharmacy students using corpus of specialized texts.

Methodology

Learner's needs

Based on the methodology described by S. Tarp (Tarp, 2008b) we attempted to establish user characteristics relevant for this dictionary project. The learner's (user) needs were assessed using on-line questionnaire. A total of 10 users (five translators working full-time with the pharmaceutical texts and five pharmacy students currently earning bachelor's degree in translation at the Institute of Linguistics and Intercultural Communication of the Sechenov University) participated in the assessment.

The questionnaire was designed to provide us with the following information:

- age, sex, and level of education (used for statistical purposes);
- primary and secondary specialty (used for statistical purposes);
- level of pharmacy-related L1 and L2 knowledge;
- most common topic of translated pharmacy texts;
- most common type of translated pharmacy texts;
- frequency of dictionary use during translation;
- most common type of terminology searched for;
- most used and preferred type of dictionary;
- frequency of using smartphone for dictionary look up;
- smartphone operating system;
- most common information that the user looks up in the dictionary;
- most common user situations.

The same questionnaire was used to address the question of optimal microstructure of the dictionary.

Thematic domains and typical user situations

Thematic dictionaries – a subtype of ideographic dictionaries – are considered to be one of the most effective tools in vocabulary building. Therefore, internal structure of the learner's pharmaceutical dictionary can be divided into several sections each of which describe relevant lexical units of a specific thematic domain. Using analysis of literature data these domains were identified by inductive method.

Results and Discussion

Learner's needs and most common user situations

The analysis of answers to the questionnaire allowed us to create user profile that describes foreseen user group. A typical user of the learner's pharmaceutical dictionary has average to above-average level of pharmacy-related L1 and L2 knowledge who will typically

translate pharmacy texts related to clinical trials, good practices (Good Clinical Practice, Good Manufacturing Practice, etc.), medicines quality control, regulatory practices and issues, herbal medicines, pharmaceutical technology, drug formulation, and medicines safety from L2 to L1.

As for the most common types of translated texts the abovementioned topics will be covered in scientific articles, clinical study reports, pharmacopoeial monographs, various technical documents, guidelines, and books. The dictionary will be consulted frequently (about one unknown word per 1-2 sentences) and the most common type of searched terminology will be either pharmaceutical or medical.

The most used (and preferred) type of the dictionary is electronic. The computer dictionary (either online or offline) will be used more often than the dictionary on smartphone (both Android-based and iOS-based). Most common information required by the user will be translation of the word and word definition and examples of sentences in which certain word is used. This is additionally supported by answers to several questions in which the responders scaled the importance of translation, definition, phonetic information, grammatical information, hyponyms/hyperonyms, ethymology, and examples from 1 (not important) to 10 (very important)

Although this profile might be somewhat subjective due to several reasons (e.g., small sample size, difference in work experience between user group and assessed group) and requires further detailing, the methodology used for its establishment can be used for obtaining preliminary information that will help to shape the overall idea of a learner's dictionary.

Optimal microstructure

Two general approaches can be used when designing optimal microstructure of a dictionary. The first was proposed by the prominent Russian lexicographer Petr Denisov: it consists of using microstructures of well-known and reputable dictionaries as a backbone for development of new dictionaries.

Using several editions of English-Russian and Russian-English dictionaries developed by I.Yu. Markovina we decided to utilize them in order to establish optimal microstructure. An example of proposed learner's pharmaceutical dictionary microstructure is shown in Figure 1.

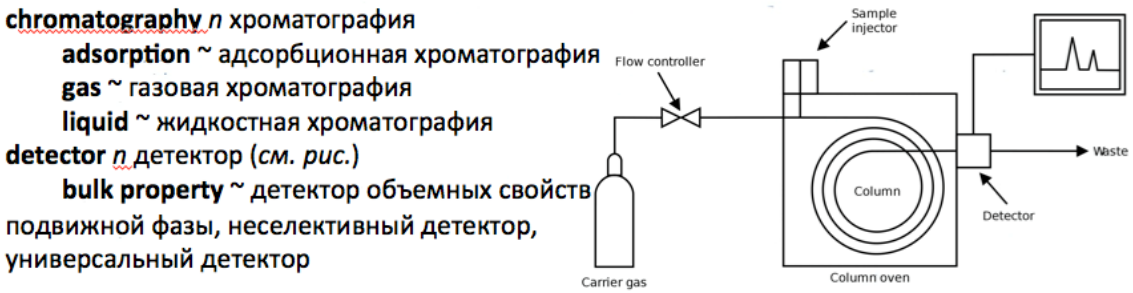


Figure 1 – Example of proposed dictionary microstructure

The lemma (shown in bold) is given in standard orthographic representation. without stress indication, syllable boundaries, and/or any other additional information. Spelling variants are show, although there will not be many of them. Phonological information is omitted since the dictionary will not be used for speech production. However, the grammatical information is in place: in our opinion it may help those learner's who struggle to determine part of speech. It is followed by semantic information (in case of bilingual LSP learner's dictionary it is the

translation equivalent). For some lemmas the label is present which contains additional information (e.g., in Figure 2 the label "*см. puc.*" points at the picture which serves as an aid for understanding and memorization of the lemmas). Most frequent collocations in which lemma is involved are present in each entry as a subentry.

Despite user requirements (see Learner's needs), in our opinion no examples of sentences in which lemma is used should be included in this type of dictionary.

The question of microstructure is essential for the optimal user experience, both in case of printed and electronic dictionary. Just like printed dictionaries are limited by physical parameters of the volume (weight, size), modern electronic dictionaries are limited by the screen resolution and its size, as well as user's health limitations (poor eyesight, etc.) However, the latter has significant advantage over its printed counterpart as it allows adjusting dictionary microstructure according to the user momentary needs. This can be achieved in two ways: during application setup by means of user dialog (e.g. "Would you like to see grammatical information by default?") or during application use by means of changing settings. Therefore, in case of electronic dictionary the question of optimal microstructure describes only how the entry will look like by default. The possibility to change background and font color, and the font itself is usually not implemented neither in on-line nor in off-line electronic dictionaries, but introduction of these options will surely improve overall user experience.

Thematic domains

A brief analysis of core competencies described in the FSES 3+ for "Pharmacy" specialty, current curricula of several Russian pharmaceutical schools, pharmacists' career profiles, relevant publications (Waterfield, 2010), and the analysis of answers to the questionnaire enabled us to produce a map of thematic domains related to pharmacy profession. A simplified version of the map is shown in Figure 2.

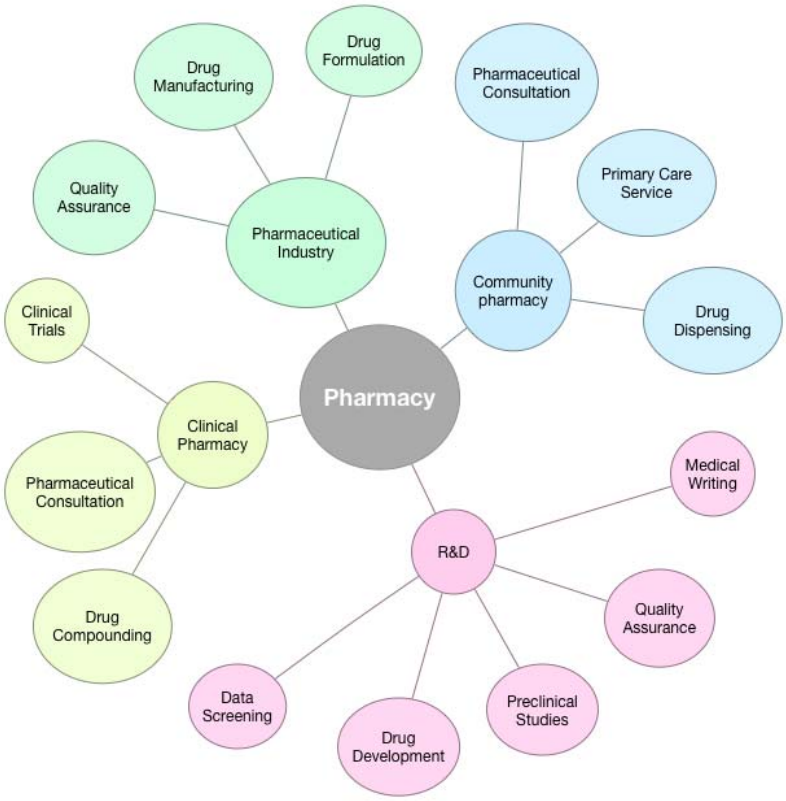


Figure 2 – Simplified map of thematic domains related to pharmacy profession

Based on the identified thematic domains we suggest that the learner's pharmaceutical dictionary can be divided thematically into four large sections and each of those can be further subdivided into several subsections as per Figure 1.

Such division can be justified by the fact that a learner's dictionary is aimed at building passive language skills: if the students use dictionary together with the similarly divided textbook they can easily find relevant information. In case the dictionary will be used in other user situations, e.g. for translation purposes, the printed dictionary should have alphabetical index. Of course, in case of the electronic dictionary there is no need for such index as the possible difficulty of use can be easily compensated by search function. An option to exclude some thematic sections from the search or from displaying can also be implemented for optimal user experience.

Further subdivision of identified thematic domains leads us to a number of fundamental scientific fields (chemistry, biology, etc.). All these fields have their own special vocabulary, which form separated lexical fields. However, these fields substantially overlap with the pharmacy lexical field to such a degree that the latter cannot function properly without this overlapping. Consider, for example, synthesis of drug substance to be used in tablet manufacturing and consequent analysis of the yielded product. The terminology of these thematic subdomains (both are included in "Pharmaceutical Industry" and "R&D" domains) is essentially the terminology of organic and analytical chemistry, respectively. Of course, terminology belonging to adjacent lexical fields should obviously be incorporated into a modern learner's pharmaceutical dictionary, otherwise the learner would struggle to find essential terms elsewhere.

Analysis of typical pharmaceutical texts shows that they contain a large number of terms frequently used in business (economics, money transfer, marketing, etc.), legal (contracts, licenses, agreements, etc.) and technical (packaging, transportation, equipment installation, etc.) contexts. In our opinion, this terminology should be included in the learner's pharmaceutical dictionary. However, care should be taken during selection process as some of the terms may appear only in very few texts and would have no relevance for the purpose of learning.

Using corpus of specialized texts to compile a bilingual LSP learner's dictionary for pharmacy students

Corpora of specialized texts are unique sources of data for dictionary compilation as they can server both as a source of information about frequency of the use of words and the databases of examples of how these words can be authentically used.

Of course, manual extraction of words from texts is always an option but in the digital world such work is time-consuming and wasteful. Different strategy should be applied when talking about the corpora of specialized pharmaceutical texts.

For example, using word list function of the freely distributed *Antconc* software developed by Laurence Anthony of Waseda University (Japan) we can obtain a list of all words occurring in the examined corpus (a compilation of pharmacopoeial monographs as they represent ultimate collection of what a pharmacist should know and, hopefully, be skilled at). The list is sorted by frequency of occurrence. A brief glance at the first 25 results urges us to move to less frequent words as the output contains, apart from articles, prepositions, and some abbreviations, words such as "standard", "system", "suitability". However, this initial opinion about words non-relevance to the field of pharmacy can be deceiving. Using concordance tool of the same software we can establish that, for example, "suitability" is encountered exclusively as part of "system suitability" and "suitability requirements". Only close inspection of relevant context and similar publications (secondary and, mostly, tertiary

sources) reveals that these two essentially mean "[chromatographic] system suitability" and "suitability requirements [for the chromatographic system]". Since chromatographic assay is the essential part of the medicines quality assessment these two words along with their collocations should be included in the dictionary.

This shows that the frequency of word occurrence *per se* is a subjective indicator of word relevance to a specialized terminology. The same lack of objectivity is observed during manual extraction of words if it is performed by a lexicographer with no experience in specific field of science. It is the combination of frequency and close inspection of relevance that does the trick. So, each extracted word and its collocations must be verified, preferably by both the lexicographer and the pharmacy professional. This can be done if data volume is of a manageable size, otherwise other methods should be employed (Schierholz, 2015).

Using specialized terminology extraction services, which usually use a combination of statistical approach, followed identification of word combinations that match certain morphological or syntactical patterns, it is possible to confirm that the right decision about word relevance was made. In the abovementioned example, one of the top results obtained by running our corpus through terminology extraction service was "chromatographic system suitability". Of course, the same critical approach to the results of such extraction must be maintained.

Taking these considerations into account, the optimal strategy for compiling learner's pharmaceutical dictionary in our opinion would be as follows:

- extraction of initial list of words from corpora by frequency;
- establishing most frequent collocations of the extracted words;
- critical examination of each word and collocation by a team of linguists and pharmacy professionals;
- additional confirmation of team's decision by terminology extraction services.

Conclusion

The aim of our paper is to share our vision of the learner's pharmaceutical dictionary concept, with particular focus on macro- and microstructure of such lexicographical work. Of course, the concept presented is subject to change as the project progresses since shortcomings and minor mistakes are almost unavoidable. Some of the steps described in the paper can be implemented when undertaking similar projects of medical and pharmaceutical dictionaries.

Detailed description of the electronic database, specific software implementation and optimal user interface are supposed to be discussed in our later publications after the completion of word extraction stage.

References

Schierholz, J. Stefan. (2015). Methods in Lexicography and Dictionary Research. *Lexikos*. 25. 10.5788/25-1-1302.

Tarp, Sven. (2008a). Lexicography in the Borderland between Knowledge and Non-knowledge. General Lexicographical Theory with particular Focus on Learner's Lexicography. Niemeyer: Tübingen. (= Lexicographica Series Maior 134). 308 p.

Tarp, Sven. (2008b). The Third Leg of Two-legged Lexicography. *Hermes, Journal of Language and Communication Studies,* 40: 117–131.

Waterfield, J. (2010). Is Pharmacy a Knowledge-Based Profession? *American Journal of Pharmaceutical Education*, *74*(3), 50.

Dictionaries cited

Bégaud, Bernard. (2000). Dictionary of Pharmacoepidemiology. West Sussex, UK: John Wiley & Sons.
González, M. Michelle (Ed.). (2008). ISPE Glossary of Pharmaceutical and Biotechnology Terminology.
Markovina, Irina Yurievna. (2014). English-Russian and Russian-English Medical Dictionary. Moscow, Russia: Zhivoj jazyk,
Nahler, Gerhard. (2013). Dictionary of Pharmaceutical Medicine. Wien, Austria: Springer-Verlag.
Philippsborn, H.E. (2007). Elsevier's Dictionary of Vitamins and Pharmacochemistry. Oxford, UK: Elsevier.
Shastri, Varun. (2005). Academic Dictionary of Pharmacy. Delhi, India: Isha Books.

A Study on the Use of the Chinese-English Dictionary: What reference skills and strategies are used by Chinese college students?

Mengyu Zhang, Lixin Xia, Chengmin Liao
Guangdong University of Foreign Studies
154472687@qq.com CDHUIYI@ALIYUN.COM

Abstract

The paper reports the results of a questionnaire survey which was recently done in China. The aim of the survey is to examine the reference skills and strategies that Chinese college students use when they consult a Chinese-English dictionary. The survey contains 8 questions, each representing a scenario in which the subjects are asked what they will do if the dictionary fails to provide productive information about equivalents. All the data from the survey were input into a database, and further analyzed with statistical tools. The survey results show that college students can adopt various strategies to cope with the situation where the active use of the equivalents lacks. The findings might be helpful for dictionary writers when they are compiling a Chinese-English dictionary.

Key words: reference skills; reference strategies; user research; L1-L2 dictionaries

1 Introduction

According to the theory of the user perspective, the reference needs and reference skills are the two most important subfields of user research (Béjoint 2000; Hartmann 1987; Svensén 2009). Although a lot of research has been done in the field of user research, the study of the skills and strategies of dictionary users is less advanced than the study of their needs (Béjoint 2000: 154) . Furthermore, the study of the L1-L2 dictionary users' strategies is still less advanced than that of L2-L1 dictionary users.

Atkins & Varantola (1997) study the dictionary look-up process of both the L1-L2 dictionary and L2-L1dictionary by the paper equivalent of the "think-aloud" protocol. They recorded each step the subjects consulted a dictionary when they were trying to translate a text either out of or into their native language. The test results show that the subjects adopted various strategies in the look-up process. For example, some Finnish participants were given two Finnish-English bilingual dictionaries, and four monolingual English dictionaries to complete their translation task. All the participants began by looking up in the bilingual dictionaries when they wanted to find an English equivalent of a Finnish expression. If they were not sure about the meaning and use of the equivalents listed in the bilingual dictionaries, they would choose to look up them in the monolingual English dictionaries provided. Apparently, they employed the orthodox, but correct strategies when seeking for a solution to a linguistic problem. Bogaards (1990) investigates the strategies employed by French speakers in their search for multi-word expression in bilingual dictionaries, and concludes that they tend to go for the least frequent word.

In foreign-language production, other strategies apply if one chooses not to use a dictionary (Rundell 1999:38): using more general expressions (e.g. very wet instead of drenched), using periphrasis (e.g. listen outside the door instead of eavesdrop), and using hyperonmys (e.g. bird instead of sparrow).

The research of the use of Chinese-English dictionary has still been a weakness in China, only few scholars have studied this field or involved Chinese-English dictionaries in their research of the use of bilingual dictionaries, they have respective emphases in their research, some of which coincide with each other while some collide. For example, many studies find that English-Chinese dictionary is more widely used than Chinese-English dictionary, while other studies are totally opposite. Moreover, when it comes to the possession of a Chinese-English dictionary, studies vary a lot. The highest record made by Yong Heming (2003b) illustrates that 73% learners have at least one Chinese-English dictionary at hand, and the lowest one found in Cui Yumei (2006) was only 9.7%. While the data remain to be investigated.

Therefore, this paper focuses on what strategies and skills Chinese college students will take when Chinese-English dictionaries offer less information about equivalents.

2 The survey

2.1 Aims

This survey, based on discrimination of equivalents, pronunciation, inflections, part of speech, grammars, collocations, pragmatical and cultural information about equivalents, aims to figure out what strategies and skills Chinese college students will take when such information is not available in existing Chinese-English dictionaries, and thus offering feedback for compilation of Chinese-English dictionaries. Different from previous surveys, this questionnaire is designed for Chinese-English dictionaries for Chinese EFL learners.

2.2 The subjects

This survey was targeted at 908 college students, among which 437 students are English majors and the rest are non-English majors. With thousands of questionnaires sent out, we received 908 sheets of answers, 803 answer sheets are valid (406 come from English major while 397 come from non-English major).

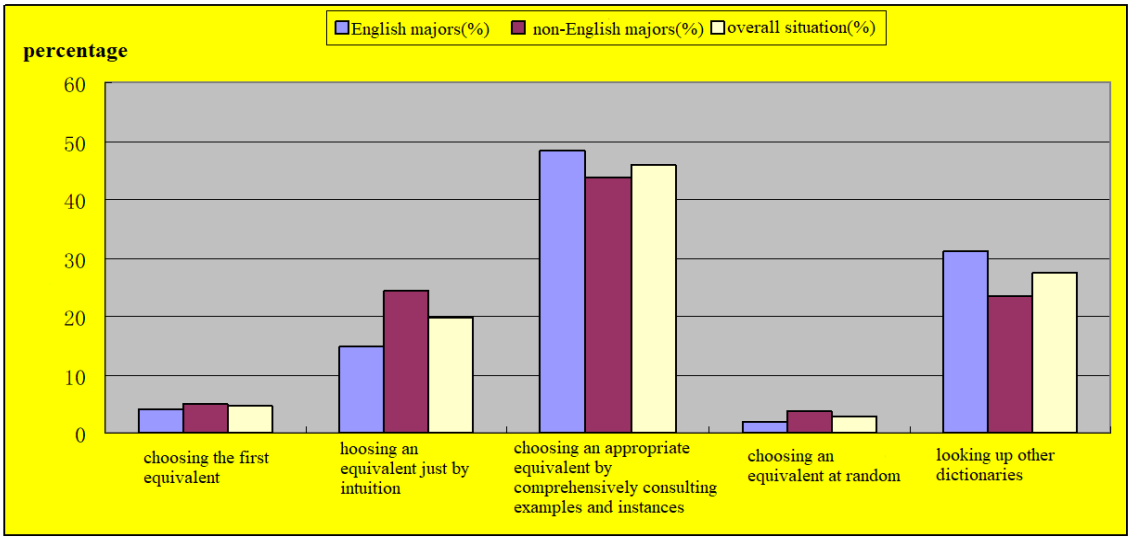
2.3 Test items

This survey was carried out in the form of structural questionnaire, subjects only choose one or more answers from given choices without filling in other information. The questionnaire has 8 questions (see appendix).

3 Results and Analyses

3.1 Selection of equivalents

Discrimination of interpretations in existing Chinese-English dictionaries still has a long way to go, users are often faced with a couple of equivalents short of discrimination or explanation. We have 5 choices here and multiple choices are allowed (see appendix). According to column 1, it is arranged by their respective proportion: choosing an appropriate equivalent by comprehensively consulting examples and instances (46%), looking up other dictionaries (27.3%), choosing an equivalent just by intuition (19.6%), choosing the first equivalent (4.5%), and choosing an equivalent at random (2.7%).



Column 1 Strategies of Chinese learners for selecting equivalents in Chinese-English dictionaries

We find that first of all, studies home and abroad (Tono 1984; Li 1998; He 2003) show that dictionary users tend to take the first equivalent. However, this survey reveals that Chinese learners seldom take the first equivalent for their language production even though they cannot tell the differences between several equivalents.

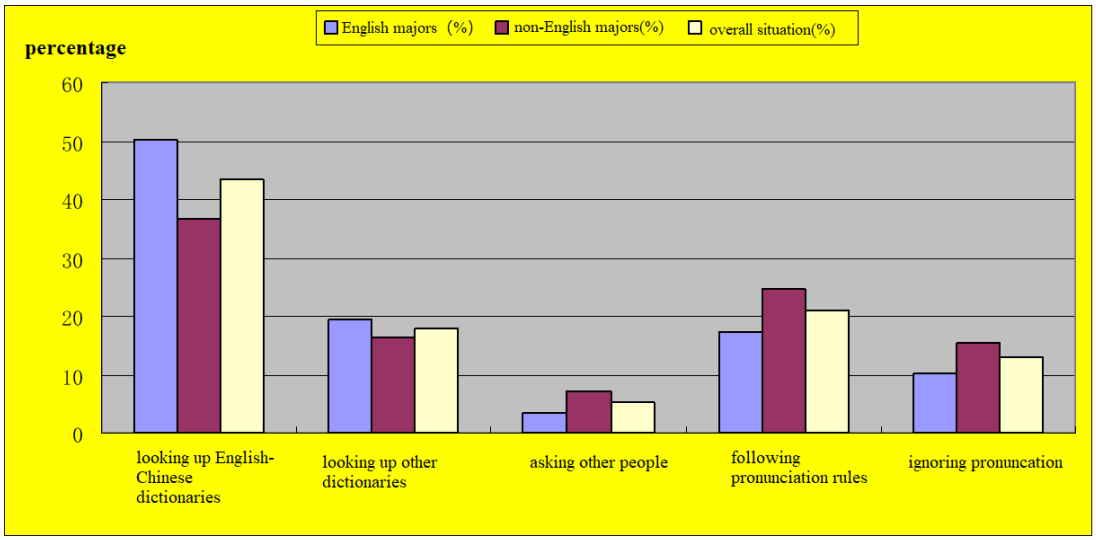
This result may have something to do with targets and methods of research. During learners' translation, influenced by original text, they will choose the first equivalent since they regard it as the needed one in their translation, this usually make sense, since modern diachronic dictionaries usually arrange equivalents in light of their importance or frequency, no matter what forms they follow, the first one is more likely to be selected for translation than the last one. But questionnaire is a different thing, subjects choose an appropriate equivalent without referring a translation task, subjectively they think the first equivalent to be unreliable therefore they have to turn to other methods.

Another finding is that selecting an equivalent by comprehensively consulting examples and instances comes to the top with absolute advantage. Before the survey, judging from lexicographic knowledge and experience of users, we predicted it to be "selecting an equivalent after looking up other dictionary". But according to this result, it is found that college students pay more attention to practicability in the meantime of pursing effectiveness.

But learners would prefer looking up words in the same dictionary to spending more time and energy on double look-up in other dictionaries. Existing dictionaries seldom provide explicit information about interpretation discrimination, users can learn the difference of sense and usage of equivalents only through examples. To some extent it reminds us that it would better meet the needs of users to solve their difficulties if Chinese-English dictionaries could offer different measures other than examples.

3.2 Query of pronunciation information

Existing Chinese-English dictionaries seldom offer pronunciation information. We surveyed learners how they deal with this problem. We have 5 questions here and multiple choices are allowed. The results are illustrated as column 2 and arranged by the proportion: looking up English-Chinese dictionaries (43.4%), following pronunciation rules (20.9%), looking up other dictionaries (17.8%), ignoring pronunciation (12.8%), and asking other people (2.7%).



Column 2 Strategies of Chinese learners for pronunciation information

Looking up English-Chinese dictionaries is mostly used by learners, which may live up to lexicographers' expectations. On the other hand we should notice that, following pronunciation rules comes to the second.

We can explain this phenomenon from two aspects. First of all, independent learning capability of learners has improved. With long-year's English courses, they basically know pronunciation rules, in most conditions they are able to pronounce correctly some words. Then, learners are loath to making double look-up in other dictionaries, they usually make self-pronunciation instead. It indirectly reflects the attitudes of users towards pronunciation information of existing dictionaries.

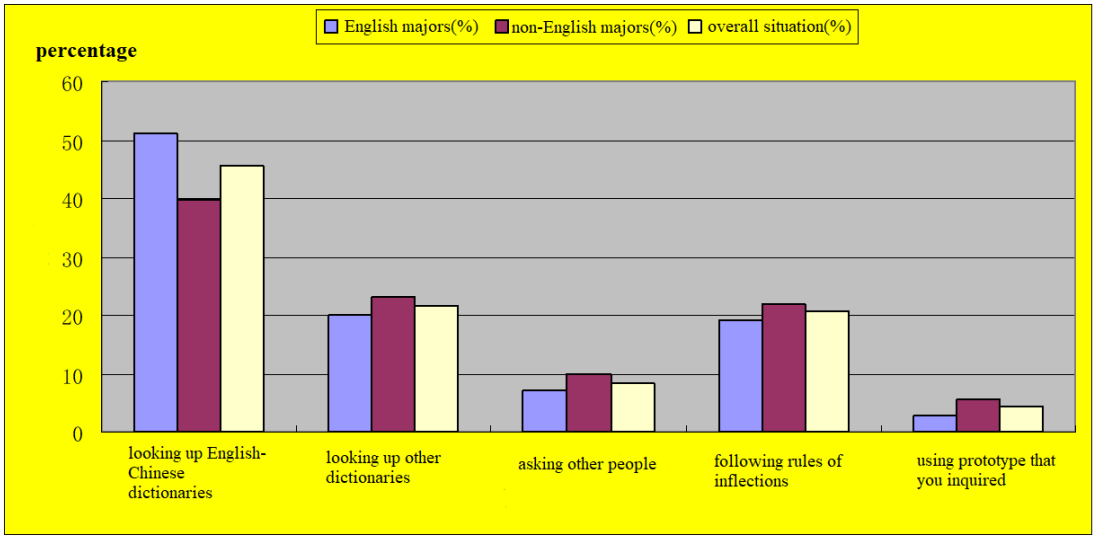
According to the data, it also shows that in lots of situations, learners would try various ways to pronounce new words, only a small amount of people (12.8%) would neglect pronunciation if they don't have to know the pronunciation.

Finally, when it comes to English majors and non-English majors, they adopt slightly different strategies. English majors tend to inquire pronunciation from English-Chinese dictionaries and other dictionaries, these two together account for 54.4%. However, in terms

of following pronunciation rules, non-English majors exceed English majors by 7% (24.5%:17.3%), it shows that non-English majors would like to independently figure out pronunciation. The reason may be that English majors attach great importance to the quality of pronunciation, they believe that authoritative dictionaries can give reliable pronunciation information though they can make it by following pronunciation rules. While non-English majors regard their goal of learning English as understanding academic literature of their majors that they have lower requirements on the quality of pronunciation.

3.3 Query of morphological information

Existing Chinese-English dictionaries do not offer inflections of English equivalents, we surveyed learners how to react to such condition and acquire right inflections of equivalents. We have 5 choices here, multiple choices are allowed. The results are illustrated as column 3 and arranged by the proportion: looking up English-Chinese dictionaries (45.4%), looking up other dictionaries (21.5%), following rules of inflections (20.6%), asking other people (8.4), and using prototype that you inquired (4.3%).



Column 3 Strategies of Chinese learners for dealing with inflections

From the above results we conclude that, first of all, learners will take various methods to get right inflections of equivalents so as to make correct production. Among these methods, learners mostly look up in English-Chinese dictionaries and other dictionaries, which together take up 66.9%. It is undoubtedly the best solution if Chinese-English dictionaries do not directly give inflections of equivalents. Learners also turn to others or try some grammar rules to work out inflections. It gives us a clue that Chinese learners will actively find a way to tackle difficulties in their process of production.

Then according to this survey, in many cases, Chinese learners solve irregular inflections with rules of regular forms of inflections, which comes to the third, constituting 20.6%, slightly lower than "looking up other dictionaries" (21.5%).

It is because learner are trying to overgeneralize information of target language. Although they have mastered some rules of word-formation, the overuse of rules gives rise to make such mistake. Therefore, before learners form a systematic knowledge about target language, Chinese-English dictionaries should predict the problems that users may encounter in coding-based language activities and take corresponding measures to help them better smooth their production.

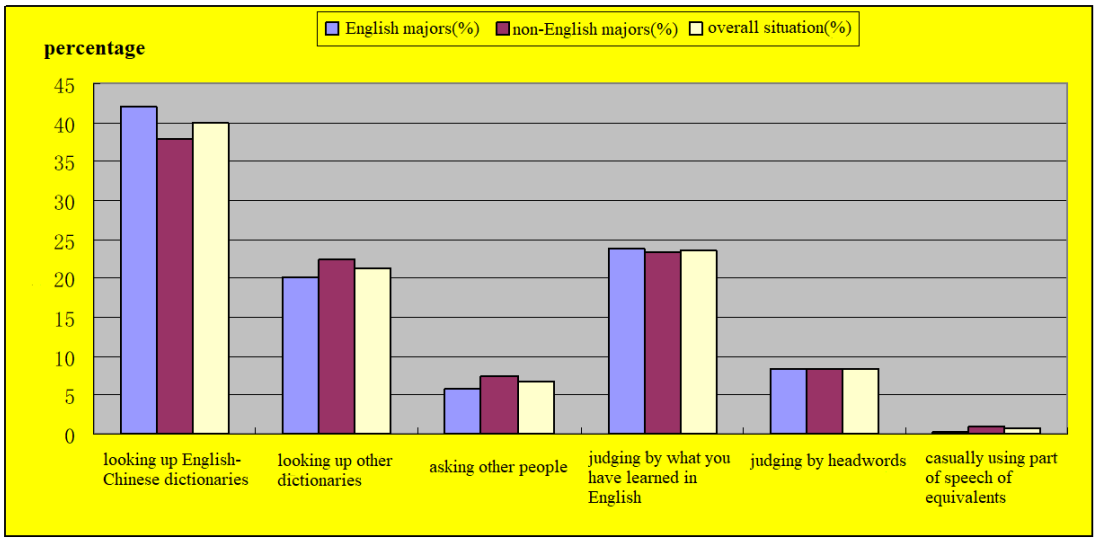
Next, using prototype that they inquired in Chinese-English (4.3%) dictionaries comes to the last, which also indirectly reflects the attitude of learners to the lack of irregular inflections in the current Chinese-English dictionary. It will free learners of double look-up

and encourage them to learn foreign language if Chinese-English dictionaries can provide alternation forms of irregular inflections.

Finally we can conclude that English majors and non-English majors take relatively same measures to deal with inflections.

3.4 Query of word class information

Current Chinese-English dictionaries only provide word class information about headwords but not equivalents, we surveyed learners how to determine word class in such condition. We have 6 choices here and multiple choices are allowed. The results are illustrated in column 4 and arranged according to the proportion: looking up English-Chinese dictionaries (39.9%), judging by what you have learned in English(23.5%), looking up other dictionaries (21.3%), judging by headwords (8.3%), asking other people (6.6%), and casually using word class of equivalents (0.6%)



Column 4 Strategies of Chinese learners for determining word class

Chinese learners mostly determine word class of equivalents by looking up English-Chinese dictionaries, if added the choice of "looking other dictionaries", like English-Chinese dictionaries (with both Chinese and English definitions) and English monolingual dictionaries, the two account for 61.2%. It shows that in many cases, learners can acquire word class information by double look-up.

On the other hand, some learners would like to tell word class with grammars they have learned, this constitutes 23.5%, coming to the second. There are two possibilities for this situation. One, learners prefer generalizing and summarizing such information with their knowledge about target language, which on the other hand may lead to errors, especially in a condition where learners can't acquire comprehensive grammars about target language.

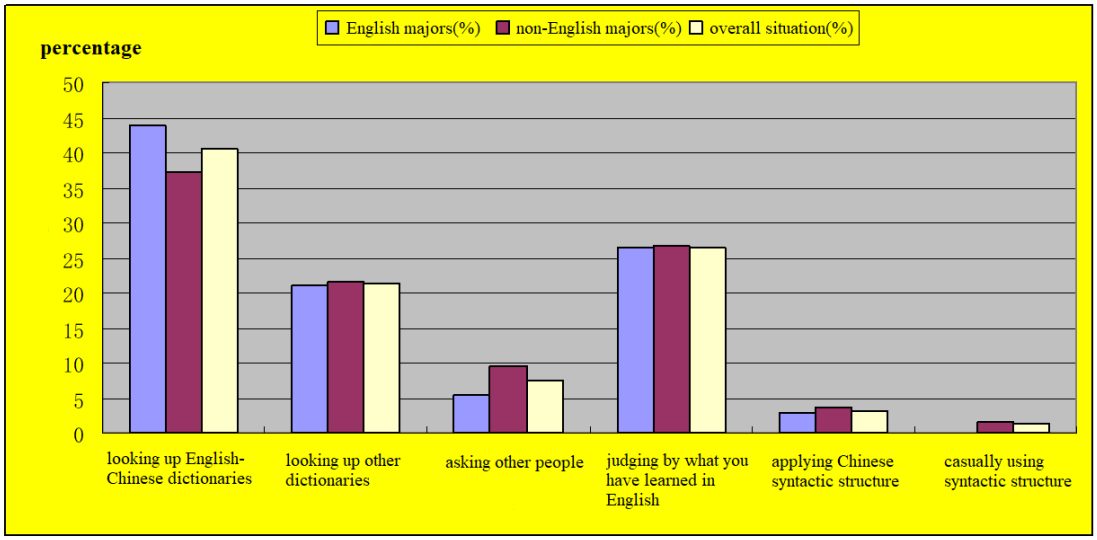
Two, since current Chinese-English dictionaries do not give any information about word class or learners would not like to look up other dictionaries or ask other people, they have to judge by themselves. No matter what situation it is, learners can be benefited if Chinese-English dictionaries can provide necessary information of word class.

According to this survey, it should be noted that learners seldom decide word class of equivalents by referring that of headwords, this choice "judging by headwords" only takes up 8.3%. In fact, recently published Chinese-English dictionaries only give word class of headwords and as much as possible use English words with same word class as equivalents, in this way learners can almost figure out word class of equivalents. However, learners may be not familiar with lexicographic compilation of Chinese-English dictionaries, they should improve their skills in using a dictionary.

3.5 Query of syntactic structure

Current Chinese-English dictionaries only provide small amount of information about syntactic structure but nothing about equivalents. We surveyed learners how to define syntactic structure of equivalents when faced with such situation. We have 6 choices here and multiple choices are allowed.

The results are illustrated as column 5 and arranged by the proportion: looking up English-Chinese dictionaries (40.6%), judging by what you have learned in English (26.5%), looking up other dictionaries (21.3%), asking other people (7.5%), applying Chinese syntactic structure (3.2%), and casually using syntactic structure (1.2%).



Column 5 Strategies of Chinese learners for determining syntactic structure

Chinese learners usually look up English-Chinese dictionaries to determine syntactic structure of equivalents. Besides, they also refer other dictionaries. As column 5 shows, their choices are relatively concentrated in these two options, which make up 61.9%, indicating that they lay much emphasis on grammar.

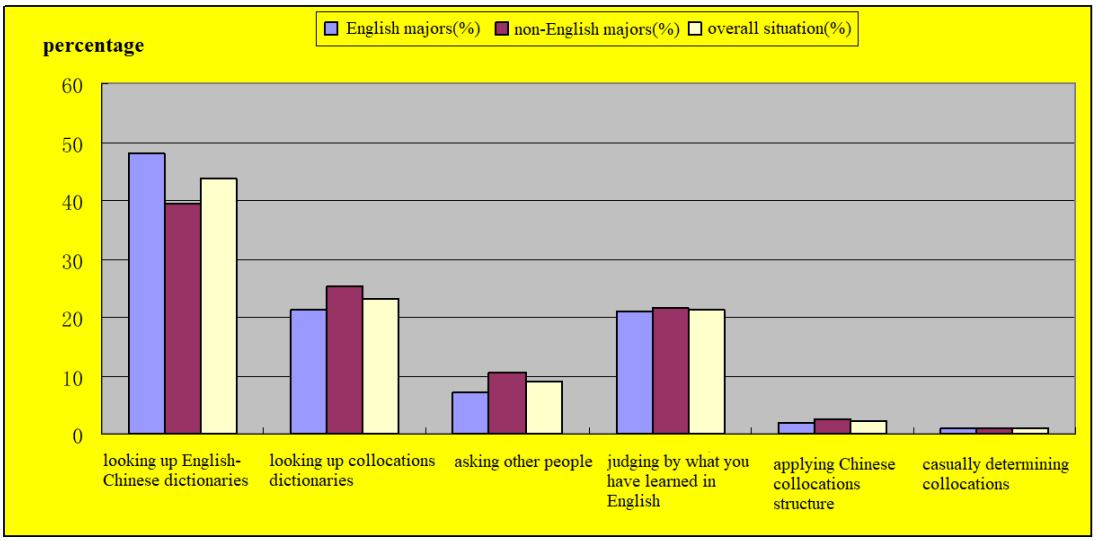
Moreover, there is a considerable number of learners trying to work out syntactic structure of equivalents with grammars they have learned, this choice ranks the second. Learners may not prefer trying other dictionaries for double look-up thus they decide to do so. To some degree, it can free them of troubles and encourage them to learn foreign language. But it may also cause disadvantageous effects.

Lastly, according to column 5 we find that English majors and non-English majors take almost the same method to handle syntactic structure, explaining that they have same demand on syntactic information.

3.6 Query of collocations information

Collocation is one of the key and difficult points for Chinese students to learn English, existing Chinese-English dictionaries explain collocations with glosses and examples, but the collocations they offered are incomplete that learners cannot find urgently needed ones. We surveyed learners how to tackle such condition and determine collocation structure of equivalents.

We have 6 choices here and multiple choices are allowed. The results are illustrated as column 6 and arranged by the proportion: looking up English-Chinese dictionaries (43.7%), looking up collocations dictionaries (23.2%), judging by what you have learned in English (21.3%), applying Chinese collocations structure (2.2%), asking other people (8.9%), and casually determining collocations(0.9%).



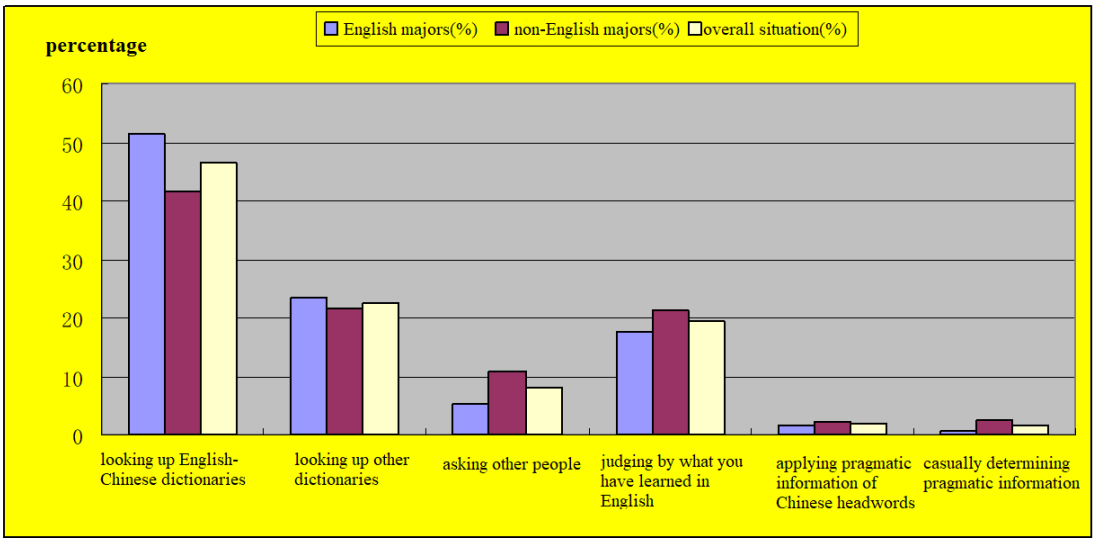
Column 6 Strategies of Chinese learners for determining collocation structure of equivalents

Similar to what discussed above, Chinese learners still take English-Chinese dictionaries as their priority, thus the reason is not repeated here.

But different from word class and syntactic information, learners would like to turn to specific collocations dictionaries to inquire collocation information, which comes to the second, accounting for 23.2%. It indicates that learners have knowledge about dictionaries and they have enhanced their skills in using a dictionary so as to work out information with different specific-aspect dictionaries. Collocations dictionaries are used for active language production, providing relatively comprehensive collocations for users, hence the likelihood of finding the required collocations will be much higher than the query in common dictionaries.

3.7 Query of pragmatic information

Current Chinese-English dictionaries often give no pragmatic information about equivalents, we surveyed learners how they deal with this condition and acquire pragmatic information. We have 6 choices here and multiple choices are allowed. The results are illustrated as column 7 and arranged by the proportion: looking up English-Chinese dictionaries (46.4%), inquiring other dictionaries (22.5%), judging by what you have learned in English (19.5%), asking other people (8.1%), applying pragmatic information of Chinese headwords (2%), and casually determining pragmatic information (1.6%).



Column 7 Strategies of Chinese learners for determining pragmatic information about equivalents

Firstly, when there is no pragmatic information of equivalents available in Chinese-English dictionaries, Chinese learners usually use English-Chinese dictionaries to solve their problems, then other dictionaries. Learners take similar measures when dealing with word class, syntactic structure and collocations, but the percentage of looking up English-Chinese dictionaries for pragmatic information is much higher. This may reflect a fact that unlike grammars and collocations, pragmatic restrictions are mainly extra-linguistic and more complicated. But luckily learners have been aware of it thus they hope to find the answer in authoritative reference books.

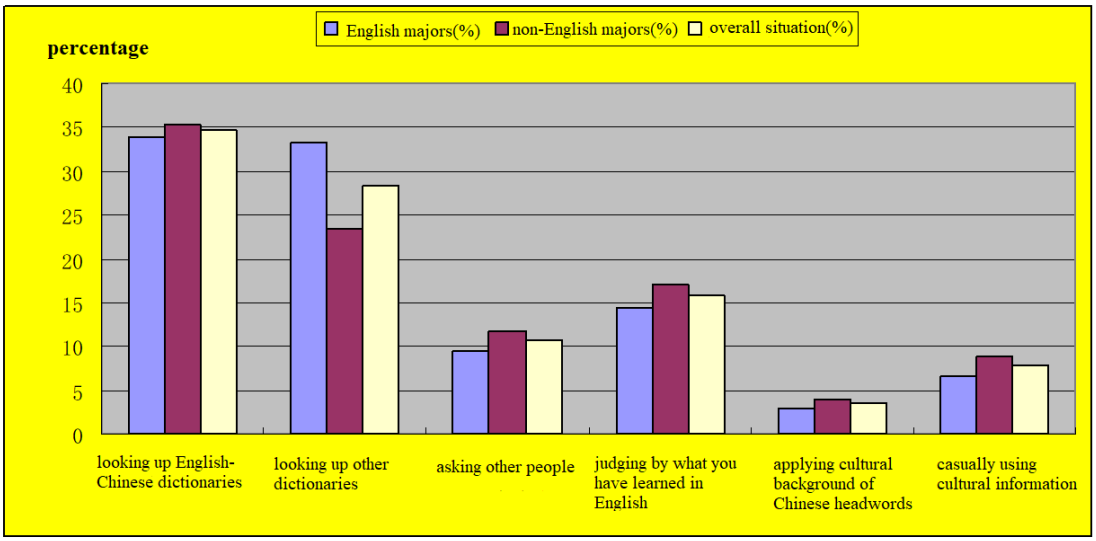
Secondly, applying pragmatic information of Chinese headwords only takes up 2%, reflecting that learners have noticed the difference of selection restrictions between Chinese and English, since selective restrictions on Chinese words do not necessarily apply to English words. Current Chinese-English dictionaries only give pragmatic information about headwords and offer English words with "seemingly similar" pragmatic information as equivalents, but in fact, they differ greatly in style, register and domain. This compilation neither meets the demand of learners for their production nor the acquisition of the second language.

Thirdly, English majors and non-English majors perform exactly opposite in the following two choices: the proportion of English majors (51.4%) who choose to look up English-Chinese dictionaries is 10% higher than that of non-English majors (41.4%); the proportion of non-English majors (21.3%) who choose to judge pragmatic information with what they have learned in English is nearly 4% higher than that of English majors (17.6%). This situation proves what we have analyzed before, with the growth of English skills of learners, they have much more requirements for their production, which can be found in their strategies for coping with pragmatic information, viz., looking up English-Chinese dictionaries is more reliable and self-judgement may lead to mistakes.

3.8 Query of cultural information

Current Chinese-English dictionaries often pay less attention to cultural effects of English words with cultural features, we surveyed the methods they usually take to solve this problem. We have 6 choices here and multiple choices are allowed.

The results are illustrated as column 8 and arranged by the proportion: looking up English-Chinese dictionaries (34.6%), looking up other dictionaries (28.3%), judging by what you have learned in English (15.7%), asking other people (10.6%), applying cultural background of Chinese headwords (3.5%), and casually using cultural information (7.7%).



Column 8 Strategies of Chinese learners for determining cultural information

On the basis of column 8, looking up other dictionaries (28.3%) ranks the second, only 6.3% lower than looking English-Chinese dictionaries. We compared the data with that of previous sections, it is found that in questions of this section, "looking up other dictionaries" ranges from 21.6% to 26.5%, in other words, there is a much higher possibility for learners to turn to dictionaries for cultural information than that for syntactic structure and collocations.

What's more, English majors and non-English majors take totally different measures to tackle cultural information of equivalents. This phenomenon may tell us that English majors think highly of cultural information that they are willing to learn cultural information by referring other related books, while non-English majors prefer using equivalents directly found in Chinese-English dictionaries.

4 Conclusions

The result of this survey basically satisfies our predicted targets, in the process of this survey, we get some beneficial findings about strategies and skills in looking up Chinese-English dictionaries, and the data can serve to study and compile Chinese-English learners' dictionaries.

According to our survey, if Chinese-English dictionaries lack necessary encoding information, lots of learners will try to find ways to solve their problems. With knowledge about dictionary popularized, skills of Chinese college students in using dictionaries improved, strategies and demands are more rational. Hence lexicographers of Chinese-English dictionary should adapt to this trend and try to present semantic information in every respect.

5 Acknowledgement

This paper was funded by the "Innovation Project of Guangdong University of Foreign Studies for Training International Postgraduate Talents".

6 Reference

Atkins, B. T. S. (1985). Monolingual and bilingual learner's dictionaries: A comparison. In Ilson R. (ed.) Dictionaries, Lexicography and Language Learning: 15-24. Oxford: Pergamon Press and the British Council.

Atkins, B. T. S. and Rundell, M. 2008. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Atkins, B. T. S. and Varantola, K. (1997). Monitoring Dictionary Use. International Journal of Lexicography, 10 (1):1-45.
Béjoint, H. (2000). Modern Lexicography: An Introduction. Oxford: Oxford University Press.
Fontenelle, T. (2008). *Practical Lexicography: A Reader. Oxford: Oxford University Press.*
Hartmann, R. R. K. and James G. (2000). Dictionary of Lexicography. Beijing: Foreign Language Teaching and Research Press.
Snell-Hornby, M. 1987. Towards A Learner's Bilingual Dictionary. In Cowie, A. P. (ed.) The Dictionary and the Language Learner :159-170. Tübingen: Niemeyer.
Snell-Hornby, M. et al (eds). (1989). Translation and Lexicography. Papers read at the EURALEX Colloquium held at Innsbruck 2-5 July, 1987: 9-20. Amsterdam: John Benjamins B. V.
Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Appendix: Questionnaire of the use of Chinese-English dictionary

College: Major: Grade:

First of all, thank you for participating in this dictionary usage survey. This survey data is only for scientific research and will not reveal any of your personal information.

1) How do you choose among synonymous equivalents if the dictionary doesn't tell you their difference?
   A. Choosing the first equivalent
   B. Choosing an equivalent just by intuition
   C. Choosing an appropriate equivalent by comprehensively consulting examples and instances
   D. Choosing an equivalent at random
   E. Looking up other dictionaries
2) What do you do if you want to know the pronunciation of equivalents in the dictionary which are a new word to you? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Following pronunciation rules
   E. Ignoring pronunciation if it is not a must
3) What do you do if you want to know the inflected forms of equivalents in the dictionary? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Following inflection rules
   E. Using prototype of equivalents regardless of infections
4) What do you do if you want to know information about parts of speech of equivalents? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Judging by what you have learned in English
   E. Judging by the part of speech of Chinese headwords

F. Casually determining part of speech

5) What do you do if you want to know the grammatical pattern of equivalents? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Judging by what you have learned in English
   E. Applying syntactic structure of Chinese headwords
   F. Casually using syntactic structure
6) What do you do if you want to know the collocation structure of equivalents? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up collocations dictionaries
   C. Asking other people
   D. Judging by what you have learned in English
   E. Applying collocation structure of Chinese headwords
   F. Casually using collocation structure
7) What do you do if you want to know the pragmatic information about equivalents? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Judging by what you have learned in English
   E. Applying pragmatic information of Chinese headwords
   F. Casually determining pragmatic information
8) What do you do if you want to know the cultural information about equivalents? (multiple choices are allowed)
   A. Looking up English-Chinese dictionaries
   B. Looking up other dictionaries
   C. Asking other people
   D. Judging by what you have learned in English
   E. Applying cultural information of Chinese headwords
   F. Using an equivalent casually regardless of cultural background

Thank you for your cooperation!

Developing a Finite State Lexicon for Sindhi

Mutee U Rahman and Hameedullah Kazi
Isra University, Hyderabad, Sindh 71000, Pakistan
muteeurahman@gmail.com, hkazi@isra.edu.pk

Abstract

Sindhi is an under-resourced language in computational linguistics and natural language processing domains. Sindhi lexical resources covering detailed morphological constructions are subject to development and evaluation. This paper describes the development of Finite State Lexicon for Sindhi by using XFST LEXC (Xerox Finite State Tools Lexicon Compiler). Developed finite state morphological lexicon covers rich linguistic details with extensive coverage of morphological forms of Sindhi word classes. Different paradigms are identified and modeled in finite state transducers using LEXC. Verbal lexicon also covers pronominal suffixation, tense, aspect and mood patterns. The developed lexicon is tested and evaluated against the corpus of 9050 words in terms of coverage, ambiguity, precision, recall and f-measure (F1). The results show 97.8% precision, 96.08% recall and average ambiguity of 1.65 solutions per word with 91.1% coverage.

Keywords: Morphological Lexicon, Language Resources, Finite State Morphology

1. Introduction

Sindhi is an Indo-Aryan language mainly spoken in Sindh province of Pakistan. Sindhi is also spoken by large populations in India and throughout the world by Sindhi immigrants (Cole, 2001). Linguistic resources for Sindhi are rarely available and Sindhi is considered an under-resourced language in the computational linguistics domain. Sindhi lexical resources covering detailed morphological constructions are subject to development and evaluation. Different natural language processing and computational linguistics applications need such resources for linguistic data processing tasks including part of speech tagging, stemming, spell checking, and syntax analysis etc. Developed finite state morphological lexicon is such type of resource which covers rich linguistic details with extensive coverage of morphological forms of different Sindhi word classes. This is an operational lexicon for Sindhi based on finite state transducers (FSTs) (Beesley & Karttunen, 2001). These transducers have upper and lower levels where the upper level represents the sequence of smaller lexical items along with their feature tags and lower level represent the surface forms produced by the upper-level sequence of lexical tokens.

In subsequent sections, existing work and implementation details are discussed in section 2 and 3 respectively. Section 4 discusses evaluation and results. Finally, the conclusion along with limitations is discussed in section 5. It may be noted that transliterated Roman script is used in the implementation.

2. Existing Work

There are a few research studies available in closely related languages like Urdu (Bogel, Butt, Hautli, & Sulger, 2007; Hussain 2004; Butt & King, 2002) and Punjabi (Virk, Humayoun, & Ranta, 2011; Humayoun & Ranta, 2010), which describe finite state morphological analyzers for these languages. However, for Sindhi, only two studies are available which include GF resource grammar for Sindhi (Oad, 2012) with a morphological analyzer and a finite state morphological analyzer for Sindhi using Apertium' lttoolbox (Motlani, Tyers, & Sharma, 2016).

The GF analyzer includes around 360 lexical entries including different part of speech classes. The study does not report the evaluation results of the morphological analysis; however, 97% accuracy is reported in terms of translation of sentences.

Apertium finite-state morphological analyzer is first ever openly available morphological analyzer for Sindhi. The coverage includes nominal and verbal morphology, the coverage in terms of number of stems is good as compared to GF analyzer (3454 versus 361) with 72 paradigms but some uncommon figures are presented; for example, 66 stems for postpositions are reported; postpositions are closed word class and in Sindhi only few postposition-stems are there. The results are evaluated in terms of precision, recall, coverage and mean ambiguity. 97.68% precision and 97.52% recall against a gold standard of 384 forms is achieved with known tokens and 97.68% precision and 72.61% recall is achieved with all tokens. The average mean ambiguity of 3.3% with 78% coverage is reported.

3. Implementing Finite State Lexicon for Sindhi

Inflectional morphology of various word classes is implemented by incorporating the inflection rules in finite-state models using Xerox LEXC (Karttunen and Beesley, 1992). Different morphological paradigms of nouns, pronouns, adjectives, adverbs, and verbs are represented through finite state transducer scripts by using LEXC syntax (Beesley and Karttunen, 2001). These scripts are compiled and transducers are generated which represent Sindhi morphological lexicon. Figure 1 shows the overall implementation model where the upper side represents the root words categorized in different POS classes and subclasses

followed by different tag sequences which represent morphological features. Whereas lower side represents surface-form lexicon or generated full form words.

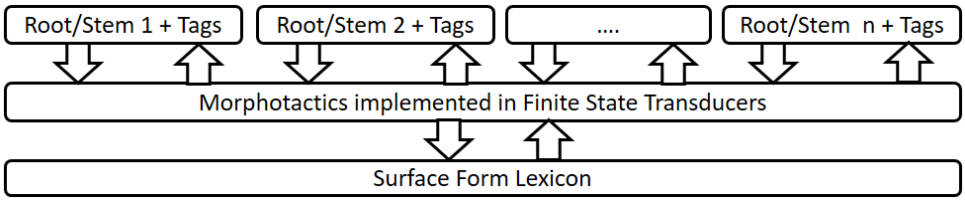


Figure 1. Overall Implementation Model

Following sections discuss the implementation details of Sindhi morphological lexicon using LEXC. It may be noted that only nouns and verbs are discussed in detail. Lexicon for the rest of the classes is implemented on same patterns.

3.1. Sindhi Nouns Lexicon

Common nouns are inflected by number gender and case. Usually, proper nouns are not inflected; however, there are exceptions of proper noun inflections in Sindhi and the inflection pattern is same as common nouns. When the number, gender, and case morphology is combined it generates up to 12 different lexical forms of a noun. Different paradigms of inflections are modeled in LEXC and resulting transducers act as function machines in which either upper side represents the input and lower side represents the output or vice versa. The reversible property of these finite state transducers (FSTs) makes them very useful. An example of a transducer is shown in figure 2. Here, when upper side is used for input, these FSTs function as surface form word generators, and when the lower side is used for input, these will function as morphological analyzers. For example, if input on the upper side is 'CHOkir+N+M+Sg+Nom' then output at the lower side will be 'CHOkirO'. Here 'CHOkir' (boy) is stem word and '+N+M+Sg+Nom' tag sequence represent noun, masculine, singular and nominative features respectively. When this information is given as input to transducer it generates the surface form 'CHOkirO' (boy). In the same way, the oblique form 'CHOkirE', the feminine nominative surface form 'CHOkirI' (girl) and the feminine oblique 'CHOkirIa' are generated. However, when this process is reversed i.e. when 'CHOkirE' is input at the lower side it will produce morphological analysis of the surface word 'CHOkirE' i.e. 'CHOkir+N+M+Sg+Obl'. Which says that 'CHOkirE' is a masculine, singular, oblique noun.

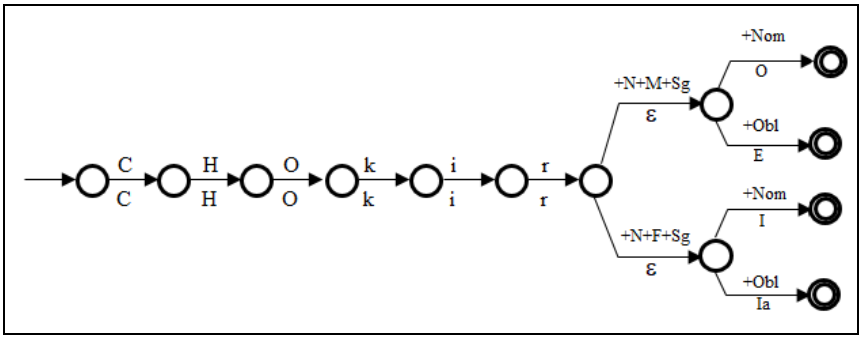


Figure 2. An example transducer showing case inflections of noun 'CHOkir'

FSTs, as given in figure2, are implemented in LEXC scripts which define finite state lexicon for Sindhi nouns. In LEXC, a Root lexicon named 'Nouns' (see figure 3) is defined which is further extended to various sub lexicons. These sub lexicons model various inflectional paradigms of nouns. Figure 3 shows the root lexicon 'Nouns' along-with 'N_Cat1' sub

lexicon. The stem form of a noun followed by a sequence of feature tags can be seen here. Stem along-with these features will produce intermediate word form shown after colon ':' following the tag sequence. This intermediate form is further inflected based on various feature sequences defined in sub lexicon 'N_Cat1'. Consider the stem form and tag sequence given below:

CHOkir+Noun+Common+Count+Animate

This will produce intermediate animate common count noun form 'CHOkir', this transducer is followed by another transducer in a series (via 'N_Cat1' sub lexicon link) which takes further input tags as shown below:

+Sg+Masc+Nominative

This produces the singular masculine nominative morpheme 'O'. The overall concatenated tag sequence preceded by the stem (upper side) and concatenated output (lower side) is given below:

Upper: CHOkir+Noun+Common+Count+Animate+Sg+Masc+Nominative
Intermediate: CHOkir O
Lower: CHOkirO

While going from upper to lower side the surface form 'CHOkirO' is generated from the stem and features specified in tag sequence; going from lower to upper will give following morphological analysis of noun 'CHOkirO'.

CHOkir {"+Noun" "+Common" "+Count" "+Animate" "+Sg" "+Masc" "+Nominative"}

Different morphological forms of stem 'CHOkir' generated by above discussed LEXC transducer are shown in table 1. The table shows case (nominative and vocative), oblique form, number and gender inflections. Total twelve (12) different inflections of the stem "CHOkir" are taken care of.

```
!SINDHI NOUN MORPHOLOGY
!AUTHOR MUTEE U RAHMAN

Multichar_Symbols
+Noun +Adjective +Adverb +Verb
+Common +Proper +Abstract !Noun Types
+Animate +Inanimate !Noun Concept
+Accusative +Dative +Ergative +Genitive +Instrumental + Locative +Nominative
+Oblique +Vocative !Noun Cases
+Count +Mass +Gerund +Measure +City +Country +FirstName +LastName +FullName +Name
+Fem +Masc !Gender
+Sg +Pl !Number
+1st +2nd +3rd !Person

LEXICON Root
Nouns;

LEXICON Nouns
        !Boy (Animate Common Noun)
        CHOkirO+Noun+Common+Count+Animate:CHOkir N_Cat1;
        .
        .
        .
LEXICON N_Cat1
        +Sg+Masc+Nominative:O                          #;
        +Sg+Masc+Oblique:E                             #;
        +Sg+Masc+Vocative:A                            #;
        +Sg+Fem+Nominative:I                           #;
        +Sg+Fem+Vocative:I                             #;
        +Pl+Masc+Vocative:aO                           #;
        +Pl+Fem+Nominative:yUN                         #;
        .
        .
        .
```

Figure 3. LEXC fragment showing part of Noun Lexicon

Table 1. LEXC generated intermediate and surface forms of noun 'CHOkir'.

Tag Sequence	**Intermediate Form**	**Surface Form**
+Sg+Masc+Nominative	CHOKir O	CHOkirO
+Sg+Masc+Oblique	CHOKir E	CHOkirE
+Sg+Masc+Vocative	CHOKir A	CHOkirA
+Sg+Fem+Nominative	CHOKir I	CHOkirI
+Sg+Fem+Vocative	CHOKir I	CHOkirI
+Sg+Fem+Oblique	CHOKir Ia	CHOkirIa
+Pl+Masc+Nominative	CHOKir A	CHOkirA
+Pl+Masc+Oblique	CHOKir ani	CHOkirani
+Pl+Masc+Vocative	CHOKir aO	CHOkiraO
+Pl+Fem+Nominative	CHOKir yUN	CHOkiryUN
+Pl+Fem+Oblique	CHOKir yani	CHOkiryani
+Pl+Fem+Vocative	CHOKir yUN	CHOkiryUN

A total of 21 different common noun categories are identified based on their inflectional properties. For every category, a different sub lexicon is defined. Most of the proper noun entries only contain feature tags without inflection paradigms. However, in Sindhi, there are exceptional cases of proper noun inflections. For example, a person name 'dOdO' can have number, and case inflections 'dOdA' (plural or vocative singular) and 'dOdE' (singular oblique form). These inflections are handled by defining another sub lexicon like common noun patterns.

3.2. Sindhi Verbs Lexicon

Verb in Sindhi is a morphologically complex word class. Verbs are marked by number, gender, case, tense, aspect, and mood. Auxiliary verbs are also inflected and marked by number, gender, and case; auxiliaries may also be used as tense and aspect markers with inflections. Copula verbs also undergo morphological changes. Due to a large number of verb categories reasonably good number of tags is used in implementation. Verb lexicon is implemented on same the patterns as nouns discussed above. Different morphological classes are defined by sub-lexicon definitions. Figure 4 shows a snapshot of few fragments of verb lexicon. Two main verbs 'likHu' (write) and 'dORi' (run) can be seen here. 'likHu' is further inflected by another sublexicon VerbStem1; after few intermediate infixations it is further inflected to infinitve, pronominal suffixes and passive endings. Infinitives, and pronominal suffix endings are also defined here, however, the definition of PassiveEnding lexicon is not

```
LEXICON Verbs
  likHu+Verb:likH       VerbStem1;!Transitive Likhu=Write
  dORi+Verb:dOR         VerbStem2;!Intransitive Type-1
  ...
LEXICON VerbStem1
  0:a                   Infinitive;
  +Psx:iy               PSuffix1;
  +Psx+PastPart+Sg:iyO  PSuffix2;
  +Psx+PastPart+Pl:iyA  PSuffix2;
  0:i                   PassiveEnding;
  ...
LEXICON Infinitive !Infinitive Formation
  +Inf:Nra              #;
  ...
LEXICON PassiveEnding
  +Passive+Sg+Masc:jE   #;
  ...
LEXICON PSuffix1
  +SSg+S1P+SMF+SObl+Sg+PastPart:ame  #;
  ...
LEXICON PSuffix2
  +SSg+S1P+SObl+SMF:mAN     OBJPSuffix;
  +SPl+S1P+SObl+SMF:sIN     OBJPSuffix;
  ...
LEXICON OBJPSuffix
  +OSg+O2P+OMF:e           #;
  ...
```

Figure 4. LEXC fragment showing parts of Verbs lexicon.

shown in figure 4. Implementation of verb lexicon covers various verbal forms, including forms shown in figure 4 (infinitives, pronominal suffixes, and passives), imperatives, causatives (four types), polite imperatives, desideratives, present participles, past participles, future participles, conjunctive participles, verbal nouns, aorists, future and passive formations along-with number and gender inflections which are not shown in figure. Pronominal suffix lexicons contain various tags to define the properties of hidden pronouns in a verb form due to pronominal suffixation. For instance, "+SSg+S1P+SMF+SObl" tags define the subject properties which say that subject is singular, first person, oblique form. Implemented transducers show that a verb can have up to 75 different lexical forms with different inflectional paradigms including pronominal suffixation.

4. Evaluation and Results

Coverage of morphology of different POS classes implemented in Xerox Finite State Transducers is shown in table 2. Coverage information shown in table 2 include the number of stem/root forms and number of morphological forms / surface forms. In this implementation, emphasis is given on the identification and coverage of inflectional categories of different word classes. Survey of morphological constructions of Sindhi through corpus analysis (Rahman, 2009) and literature review of Sindhi grammar books (Allana, 2010) provided the basis for identification of morphological patterns. As a result, 31 inflectional categories of nouns are identified and implemented along with their number, gender, and case paradigms. In the same way, 30 different categories of verbs are identified including 21 main verb categories and 9 auxiliary/copula verb categories. The number of inflectional categories and average inflections per stem for different classes are also shown in table 2. Verbs in Sindhi have most average inflections per stem. The next in this list is adjectives entry which is far behind in number (5.64 vs 53.96). Interestingly adjectives have more average inflections per stem as compared to nouns (5.64 vs 4.99); this is due to double inflections when adjectives are further inflected on adjective degree. For example, the adjective 'tHOr-O' will have all inflections (number, gender, and case) like a masculine noun with 'O' ending plus other inflections which are caused by change in degree value of adjective like 'tHor-aR-O' in this case. Now adjective form 'tHor-aR-O' will also go through number, gender and case inflections; so, if a noun with 'O' ending has 12 inflections then the adjective with 'O' ending may have up to 24 or more inflections. A total of 86 inflectional categories of different word classes are identified and implemented. The overall average inflections per stem observed are 11.8.

Table 2. Stems, Surface forms coverage, Inflectional Categories and Average Inflections

Word Class	Stems	Surface Forms	No. of Inf. Categories	Average Inflections /Stem	Overall Average Inflections
Nouns	418	2086	31	4.99	11.80
Verbs	136	7339	30	53.96	
Pronouns	129	283	7	2.19	
Adjectives	90	508	12	5.64	
Adverbs	44	45	2	1.02	
Conjunctions	18	18	1	1.00	
Postpositions	17	38	2	2.24	
Interjections	10	10	1	1.00	
Total	**862**	**10327**	**86**	**-**	**-**

The developed lexicon is tested against 9050 words corpus. Two different sets of sentences were considered. The first set contains 1390 sentences with 7809 words covering different morphological forms of various word classes. The second set contains 258 sentences with 1241 words from two textbooks of Sindhi class one. Morphological analysis is done by integrating the developed finite state transducers in XLE (Xerox Linguistic Environment) (Kaplan and Maxwell, 2016). Sample morphological analysis of a sentence is given below:

Sentence:
kAlh asIN mElO gHumaNra vayAsIN
yesterday we fair visit.Inf went.1P.Sg
We went to see/visit fair yesterday

Analysis:
kAlh {"+Adverb" "+Temp"}
asIN {"+Pron" "+1P" "+Pl" {"+Masc" | "+Fem"} "+Nominative"}
mElO {"+Noun" "+Common" "+Count" "+Inanimate" "+Masc" "+Sg" "+Nominative"}
gHumu {"+Verb" "+Inf"}
vaNGa {"+Verb" "+SPl" "+S1P" "+SMF" "+SObl" "+Pl" "+PastPart"}

In above analysis 'kAlh' is identified as a temporal adverb having +Adverb and +Temp tags, 'asIN' is a first person (+1P) plural (+Pl) pronoun (+Pron) with masculine or feminine (+Masc | +Fem) gender in nominative case (+Nominative). In the same way, 'mElO' is a common (+Common), inanimate (+Inanimate), count (+Count) noun with masculine gender, singular number and nominative case. 'gHumaNra' is an infinitive form (+Inf) of verb root 'gHumu' and 'vayAsIN' is an inflection of verb root 'vaNGa' with a plural past participle form along-with the pronominal suffix attributes starting with '+S' which represents a plural, 1st person, masculine or feminine subject in oblique form. First three words of the above sentence are root/stem words and morphological analysis only attaches the lexical attributes, however, last two words are inflected and their attributes are given accordingly.

Evaluation of the developed finite state lexicon is done in terms of precision (the fraction of words for which correct analysis is given i.e. the number of correct findings divided by total number of findings), recall (the fraction of expected correct analyses i.e. the

number of correct findings divided by the number of expected findings or gold standard), coverage (the number of morphological forms for which at least one analysis is returned), and the average ambiguity (i.e. average number of analyses returned for every word). Precision, recall, f-measure (f1), coverage and average ambiguity results are shown in table 3.

Table 3: Results of Morphological/Lexical Analysis

Word Class	Precision %	Recall %	F1 Measure %	Coverage %	Ambiguity
Nouns	97.85	96.07	96.95	96.37	1.88
Verbs	98.03	96.89	97.46	98.83	1.57
Adjectives	96.83	93.85	95.31	88.73	1.14
Adverbs	100.00	93.33	96.55	90.00	1.07
Pronouns	95.00	95.00	95.00	88.89	2.12
Postpositions	100.00	100.00	100.00	87.50	1.25
Interjections	100.00	100.00	100.00	80.00	1.00
Conjunctions	100.00	100.00	100.00	98.50	1.00
(Overall)	**97.80**	**96.08**	**96.93**	**91.1025**	**1.65**

5. Conclusion

Finite state morphological lexicon for Sindhi is presented. Di□erent inflectional paradigms of Sindhi word classes are identified and implemented in Xerox LEXC. Sindhi noun, pronoun, and adjective lexicons with number, gender, and case inflections are implemented. Sindhi verb lexicon including auxiliary, copula and modal verbs is implemented along-with their inflectional paradigms with number, gender, case, tense, aspect and mood conjugations. Usually, adverbs in Sindhi do not inflect but manner adverbs (adjectives when used as adverbs) can have number and gender inflections. Pronominal su□xation of nouns, pronouns, and verbs is also implemented with special morphological tags which reflect the information to be used for syntax analysis. Inflectional paradigms of di□erent word classes analyzed manually and modeled in LEXC transducers. Developed lexicon therefore, covers almost all possible morphological forms of an inflectional category paradigm. However, there may be rare irregular inflections in a paradigm which are not considered and as a result irregular inflections may produce some wrong results. FST lexicon is developed manually and covers su□cient paradigms which were identified for di□erent word classes. The coverage in terms of the number of entries in the lexicon is not large. However, coverage in terms of inflectional rules is quite rich and closed word classes like pronouns, auxiliary verbs, conjunctions, and postpositions are covered quite extensively. Nouns, verbs and adjective entries need to be included much more than the existing entries. Also, the lexicon developed is based on the roman transliterated script which inserts an extra layer between Sindhi text and the developed lexicon. Future work includes removing the transliteration layer and representation of lexicon in Persio-Arabic script, implementation of a guesser to guess the paradigm of a new word to be included in the lexicon automatically. This lexicon is developed and evaluated separately, however, the lexicon is also integrated with an LFG (Lexical Functional Grammar) development project. In absence of linguistic resources for Sindhi the developed lexicon will have positive impact on research and development in NLP community working on Sindhi and other South Asian language processing.

References

Allana. G. A. Sindhi Boli jo Tashrihi Grammar 'Descriptive Grammar of Sindhi Language', Sindhi Language Authority, Hyderabad Sindh, Pakistan. (2010)

Beesley, K., and L. Karttunen., Finite-state morphology: Xerox tools and techniques (preprint), The Document Company Xerox, Xerox Research Centre Europe. (2001)

Bogel, Tina, Butt Miriam, Hautli Annette, and Sulger Sebastian, Developing a finite-state morphological analyzer for Urdu and Hindi, Finite State Methods and Natural Language Processing : 86. (2007)

Butt, Miriam, and King T. H. Urdu and the Parallel Grammar project, In Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, pp. 1-3. Association for Computational Linguistics, Harvard (2002)

Cole, J. S, Sindhi, In Garry, J. and C. Rubino (eds.), Facts about the Worlds Languages: An Encyclopedia of the Worlds Major Languages, Past and Present, pp. 647-653. New York, NY: The H.W. Wilson Company. (2001)

Humayoun, Muhammad, and Aarne Ranta, Developing Punjabi Morphology, Corpus and Lexicon, In PACLIC, pp. 163-172. (2010)

Hussain, Sara, Finite-state morphological analyzer for urdu, PhD diss., National University of Computer and Emerging Sciences. (2004)

Kaplan, R., and Maxwell J., XLE Project Homepage, URL http://www2. parc. com/isl/groups/nltt/xle. Harvard. (2016)

Karttunen, Lauri, and Kenneth R. Beesley., Two-level rule compiler. Xerox Corporation. Palo Alto Research Center, (1992)

Motlani, Raveesh, Francis M. Tyers, and Dipti M. Sharma. A finite-state morphological analyser for sindhi, In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC16). (2016)

Oad, J. D. Implementing GF Resource Grammar for Sindhi language. Msc. thesis, Chalmers University of Technology, Gothenburg, Sweden. (2012)

Rahman. Mutee U, Towards Sindhi Corpus Construction, In Linguistics and Literature Review 63 Vol. 1 No 1, pp. 74-85. UMT Lahore Pakistan. (2011)

Virk, Shafqat Mumtaz, Muhammad Humayoun, and Aarne Ranta, An Open Source Punjabi Resource Grammar, In RANLP, pp. 70-76. (2011)

NEOLOGISM AND LEXICOGRAPHY: LEXICOGRAPHY CHALLENGES IN LEMMATIZING NEW WORDS IN THE SOTHO LANGUAGES

MV Mojela
Sesotho sa Leboa National Lexicography Unit
University of Limpopo
Polokwane, Rep of South Africa
Victor.mojela@ul.ac.za

Abstract:

Neologism is defined by various linguists and linguistic scholars as the creation of new words or new usage for old words, or the act of making up new words. Dictionary.com (2018) emphasizes that not all neologisms are entirely new. Some are new uses for old words, while others result from new combinations of existing words (*http://www.dictionary.com/browse/neologism)* Neologism is also known as coinage system. These newly coined words are usually absorbed into the morphological, phonological and semantical system of the language and conform to all the morpho-phonological, semantical and syntactical system of the receiving language. These new words which originate through neologism are also called adoptives, especially those words which are completely absorbed and adopted into the morphological, the phonological and the syntactical system of the language. As a result, it is important for the lexicographers to have these newly acquired terminology, systematically listed and defined for storage in the form of dictionary, i.e. lemmatized. Neologism is very much common and important in the developing languages, especially the majorities of the developing languages in the African continent like the Bantu languages in the Central and the Southern part of the African Continent. This research concentrates much on the challenges confronting lexicography in the lemmatization of the newly coined and adopted terminologies in the African languages, in particular the Sotho Languages. These Sotho languages are spoken mostly in the countries of Botswana, South Africa and Lesotho. The following are some of these challenges and linguistic complications facing lexicography in the lemmatization of adopted terminologies in the Sotho languages, i.e. ambiguities, spelling inconsistencies resulting from standardization irregularities, as well as a lack of user-friendliness of most coined lexical items which originate with neologism.

Keywords: neologism, foreign acquisitions, polysemy, coinage, standardization

Introduction

Even though the primary objective with this research is to identify and analyse the effects of neologism in the development of the African languages, it is important to mention that neologism formed the basis for the development and the increase of the vocabularies of the Bantu languages, in particular the Sotho languages. About 60% of the vocabularies of the Sotho languages originated through neologism. The Northern Sotho language which was spoken during the 19th Century is a complete different language from the Northern Sotho language spoken today due to foreign influence which were, and still are, manifested through neologism. While most of the new terms which came into being via neologism are very important to the indigenous languages, we also have instances which can be regarded to be negative developments in the languages which are directly linked with the effects of neologism. Just like most developing languages of the world, the major positive significance resulting from neologism in the South African indigenous languages is the tremendous increase of vocabulary and meaning in these languages. The increase in the vocabulary of the indigenous languages resulting from neologism also contribute very much to the development of the following negative results which can be said to be problematic to the development of the language and lexicography, i.e.

(a) widespread use of ambiguities in the languages
(b) standardization problem
(c) the creation of words which are not user-friendly, and
(d) The widening of a gap between the spoken and the written languages.

1. Background of a study

The research is based on the background studies dealing with the definition and the scholarly analysis of the term, neologism, and the analysis of both the positive and the negative effects of neologism in the Sotho languages. The works of various linguistic and lexicographic scholars such as Gouws, Prinsloo, Mojapelo and Mojela are consulted in this research.

2. Objective of the research

The main objective with this research is to give a comparative analysis of both the positive and the negative challenges which result from the use of neologism and coinage in the development of lexicography in the African languages, especially the Sotho languages in the southern part of Africa. The research concentrates much on the challenges confronting lexicography in the lemmatization of the newly coined and adopted terminologies in the Sotho Languages.

3. Methodology

In order to give intensive analysis of the consequences of neologism and coinage in the languages, a comparative methodology will be a major tool in this analytical research. Even though other subsidiary methods, like the descriptive and explanatory methodologies are also used, the comparative method is always the dominating methodology in this research.

4. Discussion

4.1. What is neologism?

Neologism is defined by many linguistic scholars as the development and adoption of new words and meanings from foreign languages through coinage. D*ictionary.com* gives the following definition of neologism:

> *A neologism is a newly coined word, expression or usage. It's also known as a coinage. Not all neologisms are entirely new. Some are new uses for old words, while others result from new combinations of existing words*

This definition is further explained as follows:

> *Some neologisms are formally accepted into mainstream language (at which point, they cease to be neologisms), and some wither until they can no longer be considered everyday terms. A neologism can be: A completely new word (e.g., over-sharers); A new combination of existing words.* **(***http://www.grammar-monster.com)*

In the Sotho languages, neologism is the system usually used to form new words or new lexical items to express the meaning of the new terms or new actions or concepts which were previously foreign to the Sotho communities. These are usually done by using the actions expressed by the foreign term or also in most cases by associating the foreign term, action or concept with the indigenous meaning. The following are Northern Sotho examples in this regard:

Aloga '*graduate*' and koma '*an initiation school'*. The term, aloga, refers to when a young man or young lady returns from a traditional Sotho initiation school, which is known as koma. This term has today undergone meaning extensions through neologism to refer also to the modern University or College graduate, e.g.:

> Koma ya bašemane e aloga gosasa. ***meaning***: *'The boys who went to the initiation school (*koma*) 'graduates' (*aloga*) tomorrow' (or, the boys come back from the initiation school tomorrow).*

As a result of neologism these terms are widely used to refer to graduating (aloga) or a graduate (se-aloga) from the modern tertiary institution. This came about as a result of the association of the old type of education with the modern type of education. Previously the word, koma, referred to the traditional type of school which was by then the only type of an institution the Sotho people used to educate their children. This term came to be associated with the modern type of education which originated with Western civilization. As a result, the words, aloga and se-aloga, have undergone meaning extension to increase their scope of meaning. The following are some of the examples of words which resulted from neologism in Northern Sotho:

Mmila '*a traditional animal path'*, but mmila today refers also to *'a road'* and 'a *highway*'

Moropa '*a traditional drum'* made from animal skin used to play traditional music. Today the word moropa refers also to a '*modern drum'* used by modern musicians'

Otlela 'drive animals by beating them'. The verb otlela comes from the verb 'otla' meaning to 'beat'. The original meaning of the word is to 'drive' or to 'beat' the animals like cattle or donkeys when pull wagons or when ploughing. With the coming of motor vehicles, the word otlela was coined to refer also to driving cars. From the term otlela originated many de-verbative nouns like mootledi 'driver' baotledi 'drivers' etc. As a result, the meanings of the traditional, otlela, extended its meaning as a result of neologism.

The example of neologism and coinage is seen in words which were formed to refer to new words by compounding, such as in words like the following:

Sellathekeng 'cellular phone'. Sellathekeng is a compound word made by combining the adverb sella 'that which cries' and thekeng 'waist'. The original cell-phones were big, and were tied on the waists. The phones literally 'cried'or gave its alarm from the waist where it was tied. Sellathekeng literally meant 'that which cries on the waist'. The following are some of the examples which originated due neologism in Northern Sotho"

Sebaledi: 'Computer'
Morethaoitiriša: 'Automatic'
Segatišamantšu 'tape recorder'

5. Neologism and its effects to lexicography development

While neologism is concerned with the creation of words and meaning in a language, lexicography is concerned with the compilation of dictionaries which serve as storage for words and their meanings in a language. While neologism is regarded as important for lexicography development since it facilitates increase in vocabulary, in some cases the lemmatization of these words is not always user-friendly to lexicography. The following are some of these negative consequences:

5.1. Neologism and ambiguity

The definition of neologism shows that new meanings are assigned to old words so that the word will now come to refer to both the old or basic meaning and the new meaning. In this case the foreign term itself is not adopted as a loan word, but only its meaning. This means that the word which originally had its own meaning is now going to undergo meaning expansion to refer to another equivalent meaning. Semanticists, like Leech (1978), have proved that no two words can have exactly the same meaning, and as a result, the extension of the meaning of an indigenous word to include the meaning of a foreign term always leads to the rise of ambiguity in the meaning of words in the languages. This ambiguity means that the words will have multiple meanings which the lexicographer should always indicate in the dictionary for the dictionary users to know. The big problem in the lemmatization of polysemous words, i.e. words which have more than one meaning, is that lexicographers are not always etymologists like the compilers of an encyclopaedia, who are required to know and to define the etymological development, or the origin, of each and every lemma in the language. The dictionary definition is different from the encyclopaedic definition of terminology where more emphasis is put on the etymological developments of words in their definition to avoid ambiguity and confusions. The Northern Sotho word, ngaka, originally referred to 'a traditional healers' and the so-called 'sangoma' in the Zulu language. The word was coined to refer also to the modern 'doctor', both the medical doctor and the academic doctors. The exact meaning of the word can only be understood when qualified in a context, e.g.:

Ke bona ngaka, will mean:

(1) '*I see a traditional healer'*
(2) 'I see a medical doctor'
(3) 'I see an academic doctor'

5.2. Neologism and Standardization

Orthography development and standardization are important for the development of languages. The newly coined lexical items are adopted to form part and parcel of the language. To be completely accepted in the language, the newly coined lexical items are standardized to conform to the orthography of the language. In most cases the lexicographers have a problem of identifying words which have already been accepted or adopted into the language and those which have not yet been adopted or standardized, especially with regard

to the acceptable spelling. In South Africa, the Standardizing Body is the Pan South African Language Board. For various reasons which include also a shortage of funds, this Body does not always meet frequently to standardize and to authenticates the newly coined words. As a result, the language usages and language developments are always moving faster than the official standardization and orthography development in the indigenous languages. Consequently, lexicographers are always faced with the problem of using words and spellings which are not yet standardized or officially accepted by the language. Consequently, the lexicographers are always having problems of lemmatizing words which are written or spelt differently. The following are examples in this regard, even though some of these words have now been standardized to conform to the correct spelling and orthography of the Northern Sotho language:

Motšhene-wa-tšhelete and Motšhenewatšhelete: 'Automatic Teller Machine' '(ATM)'

MoPresidente and Mopresidente'	'the President'
Modula-Setulo and Modulasetulo	'Chairperson"
Morwa-Motimedi and Morwamotimedi	'the Prodigal son'
Moretha-o-itiriša and Morethaoitiriša	'automatic'
Seya-le-moaya and Seyalemoya	'Radio'
Sephatša-Maru and Sephatšamaru	'Space shuttle'
Moruta-bana and Morutabana	'a Teacher'

5.3. Neologism and the challenges of user-friendliness

It is important for the dictionary to be user-friendly to the users in order to be a valuable tool for the development of the language. While the user-friendliness of a dictionary is always ascribed to various factors like the arrangements of lemma entries in a dictionary, and the accessibility of the dictionary contents, the types of the lemmas also play important role in the user-friendliness, especially with regard to the usability of words which are entered in the dictionary. Most of the coined lexical items are compound words which are formed by merging word categories such as verbs, nouns, adverbs, etc. to create new words to refer to new concepts, ideas or new meanings. These compound words are not always user-friendly to the language users, and the result is that many such words are written and adopted in the language and literature books and dictionaries are written, but are not used by people. Words like these ones are sometimes referred to as 'book terminology' (Mojapelo & Mojela, 2009) because the words are always appearing in the literature books and dictionaries but are never used by the language and dictionary users. Instead, most people prefer to use the words which were not accepted by the standard language. The people only use these words when writing because they are compelled to do so by the standardization rules, and not because they like the words. The following are examples of some of these compound words in Northern Sotho:

Paekukunama 'meat-pie', people prefer the loanword 'mitphae'
Motšhenewatšhelete 'ATM' ('automatic Teller Machine'). Preferred one is 'ATM'
Bokgobapuku. 'Library', people prefer the word 'laeporari'
Bodulabahu. 'mortuary', people use the word 'mmotšhari'
Leselawatle. 'a ship', people prefer the word 'sekepe' (from Afrikaans 'skip')
Segodišamodumo, 'loudspeaker', people prefer the word 'sepikara'
Morethaoitiriša, 'automatic' people prefer the word 'othometiki'
Dipalontshetshere, 'mathematics' people prefer the word 'mmetse' (from 'maths')

5.4. Gaps between the spoken and the written languages.

Like in most developing languages of the world, neologism is one of the major causes which leads to the creation and the widening of gaps between the written and the spoken languages, especially with regard to the use of coined loan words. This happen mostly because the coined words are usually formed and accepted by the standardizing bodies as standard words to express new concepts and new meanings. At the same time many words which develop at random without going through the standardizing bodies are not regarded to be standard forms, and are not written because the words are regarded to be inferior in status, even though these words are more frequently used than the standardized coined lexical items. This ultimately leads to the widening of gaps between the standard written languages and the spoken languages. The role of the lexicographer is to compile dictionaries by lemmatizing words which are used and written in the language even though these dictionaries may not necessarily be prescribed to be standard dictionaries. The standard dictionaries are basically regarded as dictionaries which represent the standard varieties of languages and the vocabularies contained in these types of dictionaries are standard vocabularies. Gouws & Prinsloo (2005:50) describe standard dictionaries as follows:

> '*The macrostructure primarily represents the standard variety of a treated language although a number of high usage frequency items from non-standard varieties, e.g. slang or special fields, may also be included*'

The problem confronting lexicography in this practice is that many words which have not been included as standard forms will not have uniform spelling, and different lexicographers spell the words differently in their dictionaries. This usually leads to the words being entered differently in the dictionaries since different spellings means the words will have different alphabetical order. Another big challenge to lexicography is that the dictionaries which have unstandardized and unacceptable words are not acceptable to the language authorities, who are mostly purists who regard the spoken informal varieties to be unwanted forms of the language. As such, dictionaries which include these informal terminologies are usually side-lined and sometimes prohibited by the purist authorities who don't approve these to be used by the communities and the school learners. Mojapelo & Mojela (2009) says the following in this regard:

> *As a result, the majority of the frequently used words in Northern Sotho (in accordance with Prinsloo's ruler system) do not qualify to be in the standard dictionaries, while the majority of the standard terms, which qualify for inclusion in standard dictionaries, are not frequently used words*

The following are few examples of the coined lexical items which are not frequently used in the Northern Sotho language, as compared to their variants which are regarded to be informal varieties, used only in spoken deliberations and not in formal writings:

N. Sotho (Standard)	**English Translation**	**Informal (Frequent Use)**
Sešipasameno	'toothpaste'	'Tuthpheisti'
Sefatanagantefe	'taxi'	'theksi'
Lefaseterekhomphutha	'desktop'	'deskthopo'
Go segakgopu	'overtake'	'obatheika'

6. Conclusion:

In conclusion, this research has confirmed that Neologism contributed very much to the development of the Sotho languages, especially in the borrowing and the coining of foreign lexical items to increase the vocabulary of the Sotho languages. This process has also resulted in the coining of words which are mostly ambiguous while most of these coined words are not user-friendly to both the dictionary users as well as to the lexicographers. Neologism has also led to the widening of gaps between the informal spoken languages and the written formal languages due to the fact that most coined words are accepted as standard written forms while the informal languages are usually deemed to be impure, corrupt and unwanted forms of the Sotho languages by the purists. This research gives the following recommendations as a way of reducing the challenges emanating from the improper use of neologism in the languages:

- Ambiguity is part and parcel of neologism and cannot be separated from coinage, but polysemy resulting from coinage should be clearly indicated and explained in the dictionaries by the lexicographers.
- The gap between the written and spoken vocabularies should be closed by accepting and standardizing all frequently used words in a language. Words should not be discriminated on the basis of their origin. Whether the words are derived from slang or from the inferior varieties of the language, as long as the language can convert these words to conform to its lexical and the morphological system, the words need to be standardized and accepted to form part of the vocabulary of the language. This will give lexicographers the opportunities to lemmatize words without prescriptiveness and fear of being side-lined,

References:

Allen, L. &
M.D. Linn. 1986. *Dialect and language variation.* Academic Press, Harcourt Brace Jovanovich Publishers, London, Montreal

Gouws R.H. &
D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography,* Sun Press, Stellenbosch

Dictionary.com. 2018. http://www.dictionary.com/browse/neologism
http://www.grammar-monster.com

Leech, P. 1978. *Semantics,* Penguin Books, New York

Mojapelo MW and VM Mojela (2009). *Natural Science and Technology Terminology in the Sesotho sa Leboa Monolingual Dictionary,* published in Lexikos 19 (2009)

Mojela V.M 2010. *Borrowing and Loan Words: The Lemmatizing of Newly Acquired Lexical terms in Sesotho sa Leboa,* published in Lexikos 20 (2010)

Mojela V.M. 1999. *Prestige terminology and its consequences in the development of Northern Sotho vocabulary*. Unpublished doctoral thesis, Unisa, Pretoria

Mojela V.M. 2005. *Standardization and the development of orthography in Sesotho sa Leboa – A historical overview*. *In the standardization of African Languages in South Africa,* by Vic Webb, University of Pretoria, Pretoria

A Corpus-based Analysis on Lexico-Grammatical Features in Cooking Shows

Pitchayanin Inla[1], Kornwipa Poonpon [2]*
[1] Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand;
inla.pitcha@gmail.com
[2] Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand;
korpul@kku.ac.th
* Correspondence

Abstract

Despite a large number of studies on spoken language, few have examined lexico-grammatical features in spoken discourse (Estores, 2012; Lee, 2011; Wamaitha 2014). This study investigated lexico-grammatical features which were employed in an under-researched spoken genre, cooking show. A cooking show corpus is composed of twenty video clips of cooking show hosted by one of the most famous chefs in America ranked by *www.gazettereview.com* in 2017, Gordon Ramsay. The corpus was firstly tagged, using *CLAW part-of-speech tagger,* for nineteen lexico-grammatical features, based on Bhatia (1993). Those features consisted of nine parts of speech (i.e., nouns, pronouns, verbs, adjectives, adverbs, prepositions, determiners, conjunctions, and interjections), five types of phrases (noun, verb, preposition, adjective, and adverb), and three major tenses (i.e., present, past, and future). The empirical analysis was performed, by which selected features were quantitatively determined. Occurrences of these features were calculated. The result of the study revealed that among nine types of parts of speech, verbs were the most frequently used in the cooking shows, followed by nouns and determiners. At phrasal level, noun phrases were found the most, followed by verb and prepositional phrases, respectively. Present simple were mostly found in the twenty episodes of the cooking shows, followed by future simple, present continuous, respectively. Examples of these lexico-grammatical features will be shown and discussed. The results of the study provide authentic, empirical evidence which can be a potential advantage to the field of corpus lexicography as well as a contribution to pedagogy in the field of spoken discourse.

Keywords Spoken discourse, Lexico-grammatical, Cooking shows

Introduction

Numerous studies have delved into lexico-grammatical features in different contexts. Biber and Gray (2013) investigated a lexico-grammatical analysis of writing and speaking task types on the TOELF iBT. Tseng (2011) examined the move structures and verb tense in each move from three Journals of Applied Linguistics. Also, Mehrpour and Mehrzad (2013) conducted a comparative genre analysis of English business e-mails at generic and lexico-grammatical levels. Although there is a large body of lexico-grammatical features research, most of the previous studies focused on written genre.

In the digital era in which all communication channels rely on technology, the alternative media using spoken genre, e.g., news, radio, songs, advertisements as well as television shows, seem to play an increasing role (Floyd, 2003). Despite a large number of studies on spoken language, few have examined lexico-grammatical features in spoken discourse. Estores (2012) investigated field, tenor, mode, moves and steps, as well as lexico-grammatical features in seven television talk shows. These studies are useful for audiences, television hosts, as well as ELF students in that they can improve speaking and listening skill using the language and learned structures.

Since there are various types of television shows including reality television, cooking show and soap opera (Dave, 2005). Of those shows, the cooking shows have been seen to be a preferred show because recipes and learning cooking are no longer confined to the cookbook. The cooking shows have emerged as another source of learning which provides the cookery knowledge (Ketchum, 2005; Mills, 2016; Wright & Sandlin, 2009). In addition, cooking and food have always been an area of attraction for people in society and become the topic of extensive learning in various disciplines such as medical studies, literary and cultural studies, media studies and linguistic studies (Klenová, 2010). Studies on a corpus-based analysis of cooking in spoken genre, or cooking shows, are lacking and need to be further explored. For this reason, this study aims to investigate lexico-grammatical features used in the cooking shows.

1. Background of the study

1.1 Lexico-Grammatical Features

Lexico-grammatical feature refers to the specific features of language that are used in the variety to which the text belongs e.g., tense, voice, and types of clause (Bhatia, 1993). In addition, lexico-grammatical feature is one of three under the umbrella of linguistic analysis developed by Bhatia (1993). Except lexico-grammatical feature, linguistic analysis also consists of two levels: text-patterning and structural interpretation of the text genre.

Most of research studies investigated about lexico-grammatical features in both of written and spoken genre. For instance, Saesiew (2005) analyzed the lexico-grammatical features and move structures of the motoring news in the Nation and the Bangkok post. Also, Lee (2011) studied the move structures and lexico-grammatical features employed in ESL classroom.

Lexico-grammatical features, based on Biber et al. (2004), were concerned because lexico-grammatical features are perceptive observations about the surface features, but it plays an important role to support many of the applied linguistic purposes. In addition, a text can be analyzed quantitatively by studying the specific features of language that are predominantly used in the variety to which the text belongs e.g., parts of speech, phrases, and tenses. Although these aspects are interesting and useful in term of lexico-grammatical features of various genres, it manifests very little about what aspects of genre are textualized; how successful communicative purpose is in a particular genre; and why particular linguistic features are selected.

Biber and Gray (2013) also supported that any lexico-grammatical feature that distinguishes among spoken and written will probably also be an important indicator of language development and proficiency. In this case of the study, lexico-grammatical features (i.e., parts of speech, phrases, and tenses) were focused because these features be a potential advantage to the field of corpus lexicography and help the audiences understand what the speakers want to communicate and describe in spoken genre clearer (Eastwood, 1994).

1.2 Cooking Shows

Cooking shows are a type of television genre which demonstrates what happens in the kitchen, located in the restaurants or studio sets, before the final product is presented. The purposes of the cooking shows are to help audiences learn cooking techniques and tips and also to entertain (Dave, 2005). There are several types of the cooking shows: portray and educational component (e.g., *Barefoot Contessa* and *America's Test Kitchen*), talk show (e.g., *The Chew* and *The Rachael Ray Show*) and cooking competition (e.g., *Hell's Kitchen*, and *Top chef).*

The benefits of lexico-grammatical features of cooking language are shown in some studies in the past decades. Potiantong (2010) investigated the genre of move structures and lexico-grammatical features of main dish recipes from English cookbook. Moreover, Klenová (2010) analyzed linguistics features used in cookbooks and recipes. These studies agree that analyses of cooking language are good because teachers, who teach English, can design new materials from the result of the study in order to improve students' skill. As previous studies, most of them focus on written context of cooking. Hence, more studies of lexico-grammatical features of cooking in spoken genre, or the cooking shows, seem to be worth exploring.

2. Objective

The objective of the study was to investigate lexico-grammatical features, including parts of speech, phrases, and tenses, used in a corpus of cooking shows.

3. Methodology

3.1 Cooking Shows Corpus

Cooking shows can be divided into three types (Dave, 2005): talk show, cooking competition and portray and educational component. Among these types, portray and educational component of the cooking shows were selected since a chef speaks and describes how to cook until the end of the program which makes the language can be more focused. To investigate, this paper compiled a corpus of the cooking shows. It consists of twenty episodes of the main dish cooking shows during 2017 managed by Gordon Ramsay, a well-known chef and a star in worldwide cooking programs such as *Hell's Kitchen, Kitchen Nightmares, Master Chef* as well as *Master Chef Junior.*

3.2 Lexico-Grammatical Features Analysis

CLAWS part-of-speech tagger in English was used to automatically code each lexico-grammatical feature (i.e., parts of speech, phrases and tenses) which occurs in the cooking shows corpus and *Frequency Program* was also used to calculate occurrences. As suggested from the previous studies (i.e., Potiantong, 2010; Holtz, 2011), knowledge in parts of speech helps the learners select the correct word and understand clearly what the messengers would like to communicate. Also, others previous studies (i.e., Blevins et al., 2012; Salager-Meyer, 1992; Tseng, 2011) claimed that there are different features in both phrases and tenses, concerning these features could be beneficial in order to figure out the preferred features and functions of those used in selected corpus.

4. Results and Discussion

This section reports the result of lexico-grammatical features in three areas: parts of speech, phrases, and tenses employed in the cooking shows. The analysis of lexico-grammatical features revealed that there were thirty-two lexico-grammatical features found in this corpus.

4.1 Parts of Speech

In the examination of parts of speech, the result of frequency of parts of speech in the selected cooking shows videos showed that verb is the most frequently employed followed by noun, determiner, preposition, adjective, pronoun and adverb, respectively. In contrast, conjunction and interjection were least frequently found in the cooking shows.

Table 1 Frequency of Parts of Speech in the Cooking Shows

Features in Parts of Speech	Frequency	%
Verb Normal Verb Auxiliary Verb Infinitive (*-to*) Past Participle (*-ed*) Gerund (*-ing*)	**3,284**	**21%**
Noun Material Noun Common Noun Abstract Noun Proper Noun Collective Noun	**2,923**	**19%**
Determiner	**1,935**	**13%**
Preposition	**1,879**	**12%**
Adjective Descriptive Adjective Numeral Adjective Quantitative Adjective	**1,715**	**11%**
Pronoun	**1,385**	**9%**
Adverb	**1,275**	**8%**
Conjunction Co-ordinate Conjunction Subordinate Conjunction	**740**	**5%**
Interjection	**205**	**2%**
Total	**15,341**	**100%**

As shown in Table 1, verb is mostly used in the cooking shows (3,284 times) because it is employed to describe the action or experience done by noun or subject of sentence (Alexander & Close, 1990). The normal verb of cooking process used in imperative form is highest found in this case study. For instance: *cook, start, make, chop, mix*, etc.

Noun, especially material noun and common noun, was the second most frequency of parts of speech appearing in the twenty cooking shows which is important for the chef to talk about the materials or ingredients in order to prepare the meal (Potiantong, 2010). For example, *pan, dish, oil, rice, pepper, a chuck of meat, a pinch of salt*, etc.

The following part of speech which is significantly found in the cooking shows is adjective. The finding revealed that descriptive adjectives most frequently take place because the chef uses this to describe the appearance or quality of the dish in order to grab the audience's attention and make the viewers interested in following him in cooking. Some examples of adjective used the most are *nice, good, delicious, sweet*, etc.

4.2 Phrases

The analyses so far has revealed some patterns of phrases. A phrase is a group of words which is part of sentence and categorized by considering its use with other word in the phrase (Potiantong, 2010). The result showed that the selected corpus involves 2,226 phrases. Noun phrases were the most frequently used in the cooking show corpus, followed by verb phrases, preposition phrases, adjective phrases and adverb phrases, respectively. Noun phrases were mostly found in the analysis because they were used to give specific information. For example, *olive oil, fresh basil, cold water, red wine, crunchy coleslaw*, etc. Interestingly, verb phrases were found in the second place of phrases in the cooking shows. Verb phrases always consist of the main verb and the auxiliary verb, especially the word *will,* to express the future action of preparing food. For instance, *will start, will be*, *will do,* etc.

Table 2 Frequency of Phrases in the Cooking Shows

Types of Phrase	Frequency	%
Noun Phrases	1558	70%
Verb Phrases	248	11%
Preposition Phrases	244	11%
Adjective Phrases	160	7%
Adverb Phrases	16	1%
Total	**2226**	**100%**

4.3 Tenses

The further examination in lexico-grammatical features of this corpus is tense. Tense is a significant part of the sentence because tense is the pattern of verb which describes the relevance between action and time or condition and time in order to tell when a person did something or when something happened (Alexander & Close, 1990). In this study, tense analysis concentrates on the purpose and the meaning of each tense feature toward the cooking shows. There were seven features out of the twelve features of tense found in the selected corpus. Present simple mostly found in the twenty episodes of the cooking shows, followed by future simple, present continuous, respectively.

Table 3 Frequency of Tenses in the Cooking Shows

Types of Tense	Frequency	%
Present Simple Tense	913	73.87%
Present Continuous Tense	66	5.34%
Present Perfect Tense	50	4.05%
Present Perfect Continuous Tense	-	-
Past Simple Tense	56	4.53%
Past Continuous Tense	4	0.32%
Past Perfect Tense	-	-
Past Perfect Continuous Tense	-	-
Future Simple Tense	145	11.73%
Future Continuous Tense	2	0.16%
Future Perfect Tense	-	-
Future Perfect Continuous Tense	-	-

Total	**1,236**	**100%**

The finding revealed that present simple tense was the most frequently used in order to mention something which is currently happening, describe something which occurs regularly and refer to the certain future situation. Some examples are listed as follows:

- *Boiling pasta rapidly tends to destroy the outside texture.*
- *Now they serve some beautiful baked potatoes with truffle and the salsa.*
- *It looks like large grains of rice but once they absorb the water, they double in size.*

Future simple tense is the second frequently found in the analysis. This occurred when the chef introduces the recipe of episode, describes information that possible to happen in the future or prepares the meal in the next step. For example:

- *You're gonna help me cook dinner.*
- *We're gonna now do a fig and burrata crostini.*
- *The mussels will take 4 to 5 minutes to steam.*

Present continuous tense occurs when the chef introduces the recipe he is cooking in the shows which leads to the part of demonstration and refers to the next step of cooking, as exemplified here:

- *We're creating barbecue style beef brisket with crunchy coleslaw.*
- *For my ultimate fish, I'm doing the best crunchiest lightest batter imaginable with a hint of ginger to go with it chilli minted mushy peas.*

Lexico-grammatical features, both at word and sentence levels, employed in the cooking shows have been seen clearly. The above findings of this present study provide a beneficial starting point for further studies of corpus lexicography as well as spoken discourse.

5. Conclusion

This study aimed to investigate lexico-grammatical features (e.g., parts of speech, phrases and tenses) in the cooking shows. The result from this exploration provides some benefits for people who are interested in English for cooking. The findings can make the audiences understand the process of cooking and the purpose of the speakers or chefs toward the moves and features in the cooking shows.

In terms of pedagogical implication, the lexico-grammatical features found in the cooking shows (i.e., parts of speech, phrases, and tenses) could be used to make the lesson plan for teaching English in cooking class. Moreover, lexico-grammatical found in this study could be a potential advantage to make cooking lexicography or dictionary, which is not just about the terms, but also about functions and usages. For instance, the word *cook* can be both noun and verb. These could be beneficial for beginning learners to understand cooking vocabulary and the functions of each lexico-grammatical feature used in cooking clearly. Moreover, EFL students also enhance vocabulary knowledge and improve their listening and speaking skills. They could be able to apply three main types of features (i.e., parts of speech, phrases, and tenses) correctly and appropriately.

For further studies, the samples of this study focused on the main dish cooking shows, there are other areas of interests in cooking such as bakery, dessert, or the food from different countries. The present study was conducted based on the cooking shows managed by one chef. Nowadays, there are numerous well-known chefs producing their own cooking shows. It would be interesting if further studies examine other cooking shows from various chefs in order to in order to enlarge the lexicon of this field.

6. References

Alexander, L. G., & Close, R. A. (1990). *Longman English grammar practice*. London: Longman.

Bhatia, V. (1993). *Analyzing genre: Language use in professional settings*. London: Longman.

Blevins, James P. and Ivan A. Sag. (2012). *Phrase Structure Grammar*. In M. den Dikken, ed., Cambridge Handbook of Generative Syntax. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405.

Biber, D., & Gray, B. (2013). DISCOURSE CHARACTERISTICS OF WRITING AND SPEAKING TASK TYPES ON THE TOEFL IBT® TEST: A LEXICO□GRAMMATICAL ANALYSIS. *ETS Research Report Series*, *2013*(1).

Dave, K. *Genre Analysis Reality Television and Soap Opera* [online] 2005 [cited 2017 Jul 26]. Available from: https://www.scribd.com/doc/292603230/genre-analysis

Eastwood, J. (1994). *Oxford guide to English grammar*. Oxford University Press.

Estores, R.G. *Genre Analysis of Television Talk Shows* [online] 2012 [cited 2017 Jul 26]. Available from: https://prezi.com/qcr37cgfysiz/genre-analysis-of-television-talk-shows/

Floyd, J., & Forster, L. (ed.) (2003).The Recipe Reader: Narratives, Contexts, Traditions, Aldershot: Ashgate Publishing Limited.

Holtz, M. (2011). Lexico-grammatical properties of abstracts and research articles. A corpus-based study of scientific discourse from multiple disciplines. Doctoral dissertation, Technische Universität.

Ketchum, C. (2005). The essence of cooking shows: How the food network constructs consumer fantasies. Journal of Communication Inquiry, 29(3), 217-234.

Klenová, D. (2010). The Language of Cookbooks and Recipes. Master thesis in Teaching English Language and Literature for Secondary Schools, Department of English and American Studies, Masarykova University.

Lee, J. J. (2011). A genre analysis of second language classroom discourse: Exploring the rhetorical, linguistic, and contextual dimensions of language lessons. Georgia State University.

Mehrpour, S., & Mehrzad, M. (2013).A comparative genre analysis of English business e-mails written by Iranians and native English speakers. Theory and Practice in Language Studies, 3(12), 2250.

Mills, E. We spend more time watching food on TV than we do cooking it [online] 2016[cited 2017 Jul 26]. Available from: http://www.telegraph.co.uk/food-and-drink/news/we-spend-more-time-watching-food-on-tv-than-we-do-cooking-it/

Potiantong, K. (2010). A Genre Analysis of the Main Dish Recipe. Master project in English for Business and Industry Communication, Languages of Graduated College, King Mongkut's University of Technology North Bangkok.

Saesiew, D (2005). The Genre of Motoring News in the Nation and the Bangkok Post. Master project in English for Business and Industry Communication, Languages of Graduated College, King Mongkut's University of Technology North Bangkok.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. English for Specific Purposes, 11(2), 93-113.

Tseng, F. P. (2011). Analyses of move structure and verb tense of research article abstracts in applied linguistics. International journal of English linguistics, 1(2), 27.

Wamaitha, M.L. (2014). A Genre Analysis of Argumentative Talk Shows on Selected Radio and TV Station in Kenya. Doctoral dissertation in Philosophy, English and Linguistics, Kenyatta University.

Wright, R. R., & Sandlin, J. A. (2009). You are what you eat!: Television cooking shows, consumption, and lifestyle practices as adult learning. Adult Educational Research Conference.

Integrating Thai WordNet and SenticNet into Thai Sentiment Resource

Ponrudee Netisopakul
Knowledge Engineering and Knowledge Management Research Laboratory
Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang,
Bangkok, Thailand
ponrudee@it.kmitl.ac.th

Abstract

Tremendous social network data available online has created urgent business demands to elicit insightful information in real-time. One task is to perform sentiment analysis based on text data from each online post. Sentiment analysis is a task of automatically labelling a piece of text as positive, negative or neutral, based on the sentiments of words or phrases in the text. To achieve this, a sentiment resource must be available. New pieces of text then can be automatically classified using a machine learning approach. For some standard language such as English, the sentiment resource for such tasks is publicly available. The situation is quite opposite for Thai language resource. In addition to our previous work, which constructed Thai sentiment resource using a bi-directional translation approach, this research proposes to integrate the Thai WordNet resource with the English SenticNet resource to create a new Thai sentiment resource. The new resource remedies some flaws in the previous resource. The main advantage is that it enables one Thai term to have more than one set of sentic values, based on the word senses. This new resource can be applied more realistically to the fields of opinion mining and sentiment analysis.

Keywords: Thai sentiment resource, SenticNet, WordNet, Opinion Mining, Sentiment Analysis

Introduction

Text mining research and applications have received wide interests from both public and private sectors. As text mining research advances, businesses' needs for more and more useful information are also soaring. Some traditional statistic-based approaches to text analysis such as counting word frequency and co-occurrence analysis generally do not employ semantic information. Nevertheless, these approaches have been handily applied and provided initial insights but still do not meet real businesses' needs. One immediate information a business need to know from its customers is customers' opinions toward the business's products or services. This research field is called opinion mining. In particular, the current state of the art for opinion mining is only to provide a text sentiment, that is, to label a particular opinion as positive, negative or neutral, based on the sentiment of words or phrases in the opinion text. This task is called sentiment analysis. Ideally, this information should be provided to business automatically and in real time for the business to react to the situation or to resolve the problem in a timely manner.

There are combinations of approaches to achieve this task. A pure machine learning approach would require huge manually tagged training data; while a pure semantic analysis-based approach would at least require a dictionary or a sentiment resource to begin with. A combination approach is to employ an available sentiment resource to label sentiments of words in the opinion text, then uses these words' sentiment scores to feed a machine learning algorithm to learn the overall sentiment from a set of training data. One can also directly calculate the overall sentiment of the opinion based on the words' sentiment scores. Although a machine can be employed to learn the words' importance from the independently supplied training data set, it would require humongous manually tagged opinion texts. Hence, employing a sentiment resource is more reasonable, even for machine-learning approaches. When a sentiment resource is available, a new piece of text can be automatically classified using a machine learning approach similar to the work in [1] and [2].

For some standard language such as English, the sentiment resource for such tasks is publicly available [3] [4][5][6]. For Thai language, the situation is quite opposite, Thai resources are rare. The research by [7] suggested that various Thai language resources are utmost necessary for a great deal of Thai language processing research and related fields.

In our previous research [8], we created a Thai sentiment resource using a bi-directional translation approach. A serious drawback of this approach is that, one Thai term can have only one set of sentic values and polarity. In text analysis, we often find that most terms have more than one meaning or 'sense'. Therefore, this research proposes to integrate the Thai WordNet resource [9] with the English SenticNet resource to produce a new Thai sentiment resource which enables one Thai term to have more than one set of sentic values. This new resource can be applied more realistically in fields of opinion mining and sentiment analysis.

The outline of this paper is as follows. Section 2.1 reviews Thai WordNet as a part of an Asian WordNet project. Section 2.2 reviews SenticNet resources, including their progressive methodology toward English sentiment construction and their sentiment structure. Section 3 describes our previous approach to constructing Thai sentiment resource, the so-called version 1, and summarizes its drawbacks. Section 4 proposes a new approach toward a new Thai sentiment resource, the so-called version 2. Section 5 comparatively analyzes and discusses advantages and disadvantages of both versions of the Thai sentiment resource, concluding with our plan for future work.

Background Literature

a. WordNet and Thai WordNet

Wordnet or Princeton WordNet (PWN) is an English language lexical database [10]. English words are grouped into sets of synonyms called 'synsets', each of which expresses a distinct concept and is categorized into its syntactic role as noun, verb, adjective, adverb and so on. Apart from synonyms, other conceptual-semantic relations among these synsets are established. For example, the super-subordinate relations, also called hypernymy-hyponymy relations, establish taxonomy hierarchical structure among nouns. A meronym relation specifies a part-whole relation among nouns. A troponym relation expresses more specific manners among verbs while relations among adjectives are expressed in terms of antonymy. PWN has been utilized as an initial source to create other languages' WordNet resources. An open multilingual WordNet developed by [11] has been created for more than 26 languages.

Thai WordNet was developed as part of Asian WordNet collaboration [12]. Instead of directly built Thai taxonomies, synsets and relations, which would require enormous time and budget [13], the approach described by [9] is to translate PWN's synsets to Thai using existing lexical resources. As a result, about 12% of the PWN's total synsets were able to be translated to Thai lexical entries, approximately 24,457 Thai synsets out of 207,010 PWN's synsets.

b. SenticNet

SenticNet is an English sentiment resource containing affective values of concepts, which represented as words or phrases. The affective values are arranged into four *'sentic values'* based on a theory of the hourglass of emotions [14][15]. Those four sentic values are *pleasantness, attention, sensitivity and aptitude*. According to [3], sentic values of some concepts such as *distraction* or *surprise* have negative attention values and, in contrast, positive polarity values. Therefore, an overall *polarity* value of a concept is defined as shown in equation (1).

$$Polarity(C) = \frac{Pln(C)+Att(C)-Snst(C)+Apt(C)}{9} \quad (1)$$

The SenticNet resource has been continuously developed before 2009 by a group of researchers [3][4][5][6][14][15]. Initially, the affective categorization was obtained from a Common-Sense computing project call 'The Open Mind Common Sense project'. The project website has been online since 2000 to collect Common Sense knowledge from volunteers in order to provide intuitions to AI applications. One result of this project is an ontology called *ConceptNet*, represented as a directed graph in which the nodes are concepts and the labeled edges are assertions of Common Sense interconnecting them. Figure 1 is an illustrative sketch of ConceptNet as a directed graph.

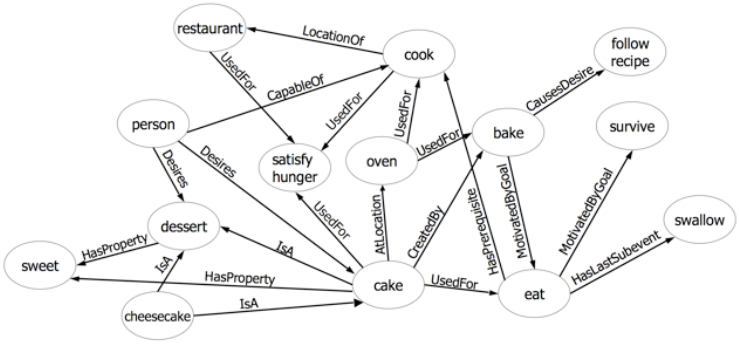


Figure 1. A directed graph in ConceptNet [15]

Source knowledge is a WordNet-Affect, developed from WordNet by identifying synsets marked with a label `emotion', 'mood', or a situation eliciting emotions. In order to

perform emotive reasoning utilizing information from these two resources, a *blending* technique was developed to combine two knowledge sources into one humongous multidimensional vector space representation called *AffectiveSpace*. In this space, the lemma forms of the words in WordNet-Affect were aligned to the lemma forms of concepts in ConceptNet. Dimensional reduction using Singular Vector Decomposition (SVD) was then performed on this huge matrix to obtain the prominent 50 to 100 principle components, in other words, 50-100 dimensions in the AffectiveSpace [3][4]. SenticNet 1 and SenticNet 2 were developed along mainly with the described methodology. In addition, a repetitive operation called *'spectral association'* was applied to help spreading affective values across the entire ConceptNet graph. Another Concept-Frequency Inverse-Opinion-Frequency or CF-IOF weighting method was also applied to evaluate importance of a concept to a specific topic. The representation of SenticNet 2 was generally improved over SenticNet 1 in many aspects. The resource was encoded in RDF/XML format using the descriptors defined by Human Emotion Ontology (HEO); each term was associated with four sentic values, its polarity and top ten similar terms. According to [4], this was apt for social data mining.

Table 1. Comparison of four versions of the SenticNet resource

Vesion	Published Year	Sources	Affective Result(s)	Final resource representation	Number of terms (Approx.)
SenticNet 1	2010	ConceptNet from Open Mind Common Sense (OMCS) corpus and WordNet-Affect	AffectNet	Term and Term's Polarity value	5,700
SenticNet 2	2012	ConceptNet and Live Journal	AffectiveSpace	Top ten similar concepts Four sentic values	14,000
SenticNet 3	2014	COGBASE, Cyc, ConceptNet, WordNet, YAGO	AffectiveSpace allows for multiword expressions	Associated semantic concepts Four sentic values Polarity value	25,000
Senticnet 4	2016	Noun and Verb Concept Generalization on previous SenticNet	Generalization of AffectNet and	Sentic patterns to infer polarity	50,000

SenticNet 3 employed an additional resource called the COGBASE commonsense knowledge formalism. This resource contained 30,000 multiword expressions, which "enable a deeper and more multi-faceted analysis of natural language opinions" [5]. SenticNet 4 extended and generalized the previous SenticNet by linking verb and noun concepts to their primitives. The related primitive concept discovering was performed automatically using a hierarchical clustering algorithm and a dimensionality reduction utilizing VerbNet and SenticNet LDA. When performing experimental comparisons on the sentence-level Movie Review Dataset, SenticNet 4 gained better accuracy than the previous SenticNet [6]. The summarization of the four SenticNet versions is shown in Table 1.

Previous Thai Sentiment Resource: Bi-directional Translation Approach [8]

The bidirectional translation approach toward constructing Thai sentiment resource utilized a publicly available English sentiment resource called 'SenticNet' version two [4] as a starting point. This approach then translated English terms in SenticNet into Thai terms using two sources of English-Thai dictionaries, LEXiTRON [16] [17] and Volubilis [18]. One term in SenticNet2 can be translated into many Thai terms and vice versa. Only those Thai terms that can be translated forth and back into the original English term are kept. Any Thai terms obtained from the English-Thai forward translation that cannot be translated back

from Thai to the original English terms are discarded, hence, deriving the approach's name 'bi-directional translation'.

In this approach, when one Thai term is bi-directional translated to many English terms, the Thai sentic values are calculated as averages of those English sentic values. Four sentic values are calculated, namely, Pleasantness, Attention, Sensitivity and Aptitude. Therefore, one Thai term eventually has only one set of sentic values. Finally, Thai terms and their sentic values are stored into Thai sentiment database. The processing steps of this approach is shown in Figure 2.

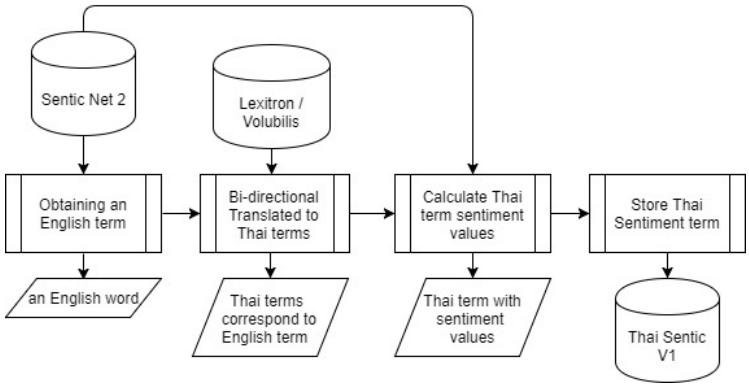


Figure 2. The Bi-directional approach to constructing Thai sentiment resource

Out of 14,244 English terms in SenticNet2, this approach impressively obtained 16,584 Thai sentiment terms using both dictionaries. The resource was utilized to extract prominent polarity terms from 1,964 sentences of 40 Thai children stories. These sentences were manually labelled as training data and SVM machine was applied to classify their sentiments. It achieved about 72% accuracy compared with human's [1]. A further in-depth analysis was carried out to identify the cause of the sentiment misclassification [2]. There were mainly two causes related to the Thai sentiment resource. First, the prominent terms chosen by human were not in the resource, hence, these terms were not chosen by the machine; consequently, the classification was not accurate. Second, meanings or senses of those Thai terms, which could be translated into many English terms, were amalgamated into only one sentiment polarity. Therefore, the sentiment polarity chosen for a particular sentence was wrong, causing the inaccurate classification. In short, the Thai sentiment resource version 1 needs to be improved.

Proposed Thai Sentiment Resource: Thai WordNet Integration Approach

It is commonly known that one word or one term can have many meanings or senses. A WordNet ontology is a language resource that contains many different meanings or senses for one term, one sense per one 'synset'. Thai WordNet is a part of Asian WordNet [9] [12]. Our new approach to constructing Thai sentiment resource relies on Thai WordNet to obtain a term's many senses. The processing steps go as follow. First, Thai term is fed from a dictionary one at a time. Two sources of the dictionary are utilized here; those are LEXiTRON and Volubilis. The term is then matched with the same term in Thai WordNet to obtain its synsets. Note that synsets are presented as English meanings of the Thai term. If the term is not found in Thai WordNet, then it is ignored. Now each Thai term is associated to possibly one or many Thai-English synset pairs. Each pair will be assigned a set of sentiment values based on the SenticNet 4 resource [6]. Any English term does not appear in the SenticNet 4 resource, then the sentic values cannot be assigned, therefore, the synset pair will be ignored. Figure 3 depicts the processing steps of constructing Thai sentiment resource based on Thai WordNet.

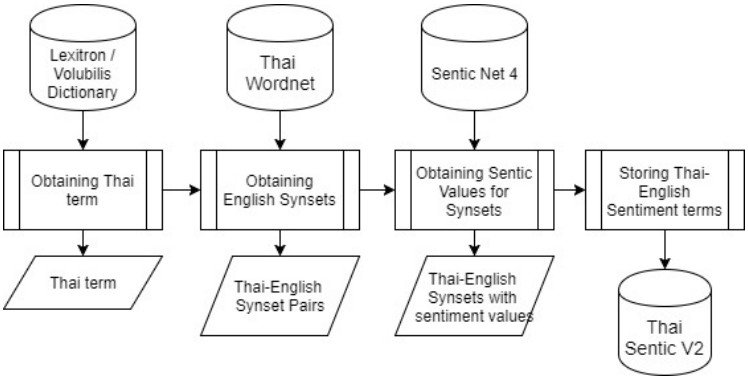


Figure 3. Thai WordNet approach to constructing Thai sentiment resource

proposed approach obtained 43,156 synsets with 30,633 distinct Thai terms, about two times larger than the previous version. For reference information, the combined LEXiTRON and Volubilis dictionary has 186,228 terms while Thai WordNet has approximately 91,073 synsets.

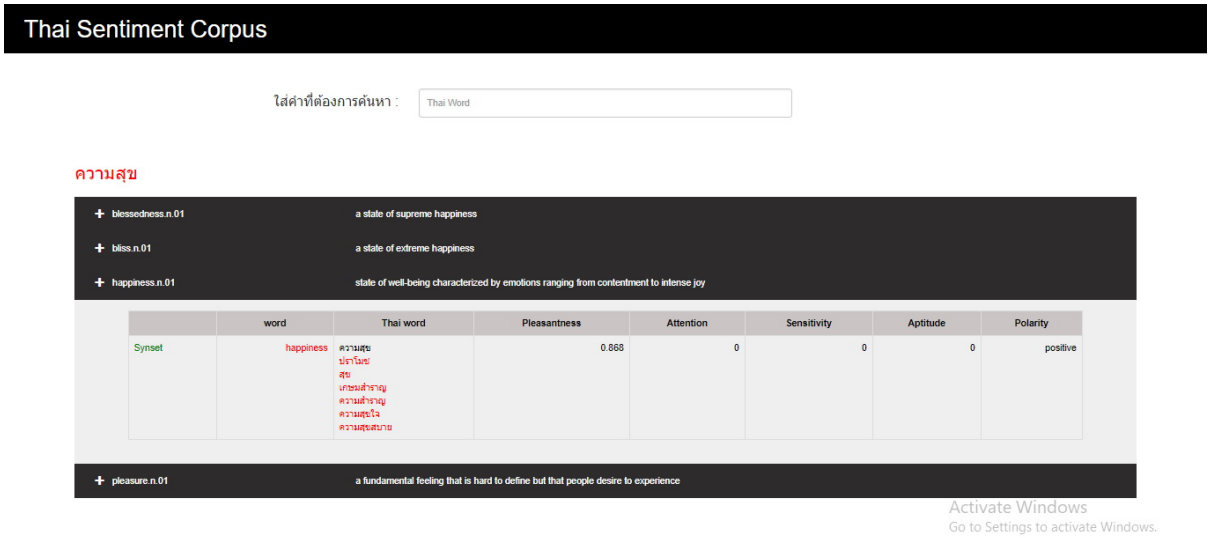


Figure 4. The web-based browser for Thai sentiment resource version 2

A web-based browser for Thai sentiment resource version 2 is developed to easily access to the resource. Figure 4 demonstrates resulting synsets for a word 'ความสุข', which is mapped into four synsets: blessedness, bless, happiness and pleasure. Each synset can be expanded to access to associated Thai terms. These expanded terms also have a hyperlink to their own synsets. Visually, this process can be resulted in a network of connected synsets, as shown in Figure 5.

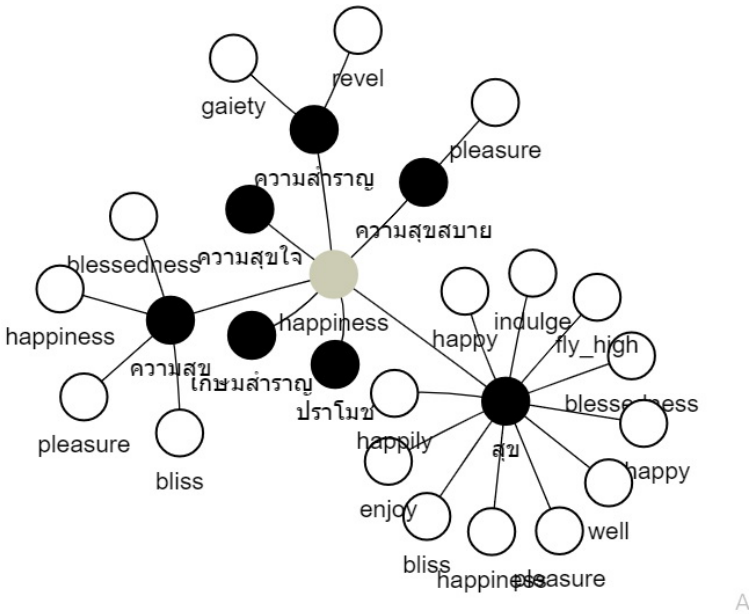


Figure 5. Related terms correspond to a word 'happiness' from Thai Sentiment resource Version 2

In Figure 5, the center grey dot represents a synset 'happiness' that links to other seven Thai terms shown in black dots. Those are 'ความสำราญ', 'ความสุขใจ', 'ความสุขสบาย', 'ความสุข', 'เกษมสำราญ', 'ปราโมช' and 'สุข'. Some of these terms also link to other synsets shown as white dots in the outer ring, resulting in an interlinked network of 'happiness'. Beneficially, this synset-based approach combined with the browser capability can enable visualization of interlinked network for related terms.

Comparison of Two Thai Sentiment Resources and Future Work

Table 2 compares the two versions of Thai sentiment resources. In the second version, Thai WordNet and more updated SenticNet resource are added to the source data. The second version has several advantages over the first one. First, the resulting number of distinct Thai terms increases from about 16K to 30K, almost doubled. The second and also main advantage of the second version is that it differentiates different meanings of a same term through synset senses. This enables more realistic application of sentiment labeling in different domains. Another side advantage but not less significant is that there is a web-based browser implementation to search Thai term's synsets, their associated sentic values and polarities. These synsets in turn connect to other related terms, enabling exploration of the term's network.

Table 2. Comparison of two versions of Thai sentiment resources

	Thai Sentiment Resource	
Source data	Version 1	Version 2
	SenticNet 2	LEXiTron,Volubilis dictionary
	LEXiTron,Volubilis dictionary	Thai Wordnet
		SenticNet 4
Approach	Bi-directional Translation	Thai Wordnet Synsets
Number of terms	16,584 Thai terms	30,633 distinct Thai terms
Number of senses	16,584 (One sense per one term)	43,156 Thai-English Synsets
Available	A resource text file	Text fiel and a web-based brower
Deployment and Evaluation	Sentence-level sentiment prediction on Thai children stories	(Plan) Sentiment classification on Social network text

More deeper investigation reveals that there are 23,115 terms in the second version that the first version does not have; while there are 8,509 terms that appear in the first version only but not in the second version. The most likely cause of more terms in the second version is the more coverage of SenticNet 4 over SenticNet 2. The most probable cause of those missing terms in the second version is the loss while either passing through Thai WordNet synsets or through SenticNet 4. This requires further investigation. In addition, there are only 7,518 terms in common contained in both versions. Therefore, to gain most benefit, the 8,509 terms missing from the first version can be augmented to the second one.

An investigation with terms in the second version, which have many senses, also reveals 2,263 terms holding opposite polarity values. An example of these terms is "กลมกลืน", translated into *shading* with a negative polarity, *match, harmonize and harmony* with positive polarity. This is an acceptable case. However, there are also some curious cases, such as the word "เมตตา", which can be translated to *mercifully, and mercifulness* - with positive polarities and also can be curiously translated to *unkind* with a negative polarity. This kind of terms is needed to be further investigated.

A future plan is to apply this new Thai sentiment resource to a social network data, such as posted text from Facebook, Twitter, or Thai social network websites, such as pantip.com.

References

[1] Lertsuksakda, R., Pasupa, K., and Netisopakul, P. (2015). "Sentiment analysis of Thai children stories on support vector machine" In Artificial Life and Robotics (AROB), 2015. Proceeding of the Twentieth International Symposium on. Beppu, Japan, pp. 138-142.

[2] Netisopakul, P., Pasupa, K., and Lertsuksakda, R. (2017). "Hypothesis testing based on observation from Thai sentiment classification" Artificial Life and Robotics, 22(2), 184-190.

[3] Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010, November). Senticnet: A publicly available semantic resource for opinion mining. In AAAI fall symposium: commonsense knowledge (Vol. 10, No. 0).

[4] Cambria, E., Havasi, C., and Hussain, A. (2012). "SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis," In FLAIRS conference, May. 2012, pp. 202-207.

[5] Cambria, E., Olsher, D., & Rajagopal, D. (2014, June). SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In Twenty-eighth AAAI conference on artificial intelligence.

[6] Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2666-2677).

[7] Netisopakul, P. and Wohlgenannt, G., (2018). "A Survey of Thai Knowledge Extraction for the Semantic Web Research and Tools" IEICE TRANS. INF. & SYST., vol. E101–D, no. 4, 2018.

[8] Lertsuksakda, R., Netisopakul, P., and Pasupa, K. (2014). "Thai sentiment terms construction using the Hourglass of Emotions" In Knowledge and Smart Technology (KST), 2014 6th International Conference on (pp. 46-50). IEEE.

[9] Thoongsup, S., Robkop, K., Mokarat, C., Sinthurahat, T., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2009, August). Thai WordNet construction. In Proceedings of the 7th workshop on Asian language resources (pp. 139-144). Association for Computational Linguistics.

[10] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[11] Bond, F., & Foster, R. (2013). Linking and extending an open multilingual wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1352-1362).

[12] Sornlertlamvanich, V., (2016). Asian WordNet Framework: Its Web Service and Collaborative Platform. The Second Wordnet Bahasa Workshop, NTU, Singapore. January 15-16, 2016.

[13] Leenoi, D., Supnithi, T., & Aroonmanakun, W. (2008). Building a gold standard for Thai WordNet. In Proceeding of The International Conference on Asian Language Processing 2008 (IALP2008) (pp. 78-82). COLIPS.

[14] Cambria, E., Livingstone, A., & Hussain, A. (2012). "The hourglass of emotions." In Cognitive behavioral systems (pp. 144-157). Springer, Berlin, Heidelberg.

[15] Cambria, E., Hussain, A., Havasi, C., & Eckl, C. (2009). AffectiveSpace: Blending common sense and affective knowledge to perform emotive reasoning. WOMSA at CAEPIA, Seville, 32-41.

[16] LEXiTRON Team, "Thai-English Dictionary – LEXiTRON" (2009) [Online] http://LEXiTRON.nectec.or.th/2009_1/

[17] Jumpathong, S., Phaholphinyo, S., Leenoi, D., Kriengket, K., Boonkwan, P., & Supnithi, T. (2016). LEX-iTRON Dictionary Improvement based on the Computational Lexicography Method. The new Journal of Intelligent Informatics and Smart Technology (JIIST), 1.

[18] Belisan, "Volubilis dictionary" (2013) [Online] http://belisan-Volubilis.blogspot.be/.

Indigenious Languages and Learner Dictionaries: The Use of Javanese Dictionary in University

Totok Suhardijanto
Universitas Indonesia
suhardiyanto@gmail.com

Atin Fitriana
Universitas Indonesia
atinfitriana@gmail.com

Abstract

It is widely known that Indonesia is one of the most linguistically diverse countries in the world. Some major Indonesian universities has been offering a degree program on several major indigenious languages such as Javanese, Sundanese, Balinese, Malay, Bataknese, and Buginese.

This paper presents our studies on the use of Javanese learner dictionary in University of Indonesia. The Javanese language is chosen for three reasons. First, in terms of number of speaker, Javanese is considered as the main indigenous language in Indonesia. It is spoken by more than 84 million people (Ethnologue 2017). Second, in terms of documentation, Javanese has a long writing tradition in Indonesia, which in turn will have wider and deeper influences to its people. Third, Javanese is also taught as local contents from elementary school to high school in East and Central Java, as well as a northern part of West Java.

The problem is whether dictionary as a supporting instrument of learning is working properly or not? In some major university in Indonesia where Javanese is taught either as a major or as a minor program, not all learners are native speakers. For those who are admitted to a university program that offers a Javanese language and culture, some of them are even just starting to learn Javanese language at the university level. According to some learning participants, the difficulty is mainly in terms of words selection, due to the absence of examples of synonymous word usage and the absence of a context that explains what and how it differs. In addition, Javanese language has many dialects and registers. In this matter, the native speakers are facing difficulties.

As a result, in this paper, the issues is how far the preparation of lexicographic works that can help non-native and native speakers to overcome these gaps? How is the presentation of the entry, gloss, and the definition so that the provided dictionary meets the needs of non-native and native participants?

Keywords: Javanese language, lexicography, learner dictionary

Introduction

A dictionary is a reference book in an alphabetic arrangement where all information about lexicons of a language can be found. These informations comprise of word forms, pronunciations, functions, meanings, etymologies, spellings, variation, idiomatic uses, and examples of a particular word use in a context. Most of this information cannot be found in other reference books. For this reason, a dictionary is one of the most crucial stuffs for a foreign language learner, especially as it makes the learner more independent of the teacher.

This paper presents our attempt to create a new Javanese learner dictionary to meet the student needs in university level. Javenese is the most important language in Indonesia with the largest number of speaker. Under Indonesian law and constitution, languages in Indonesia are classified into three different categories: national, local/indigenious, and foreign languages. Within this scheme, Javanese is regarded as a local or indigenious language. According to Language Act, Indonesian indigenious languages function as (i) communicative means for local goverments in a particular time and condition; (ii) instruction languages in elementary schools; (iii)media languages in local mass media; (iv). ritual languages in local and religious activities; and (v) geographic names of administrative regions, buildings, road, and organization.

Javanese is spoken by 84,3 million people in three provinces in Java, including Yogyakarta Special Region. It has three main dialects, that is eastern, central, and western dialects, and three different styles or registers: ngoko, madya, and krama. In addition to Malay, Javanese is also one of the most studied local languages in Indonesia. The language becomes the language of instruction in elementary education in several provinces of Indonesia, mainly in Java island.

Being so widespread, Javanese has a number of dialects. Wedhawati, et al. (2006) mentions there are at least 3 Javanese dialect, namely Yogya-Solo dialect (considered as the standard dialect), Banyumas dialect, and East Javanese dialect. Further, each dialect has its own subdialects (see Poerwadarmina 1953; Uhlenbeck 1964). Generally, the distinctions between one dialect to another lie on the phonological, morphological, and lexical aspects.

In addition to the number of dialects, Javanese is also considered as one of the difficult languages in Indonesia due to its character and registers. Similar to any major local languages in Indonesia, Javanese adopted a Brahmic script and used it over centuries as own script named *Hanacaraka script* until prohibited during Japanese occupation in the early '40s. In recent days Javanese people uses a latin script foremost in many occasions although Javanese script is still taught in elementary schools.

With regard to Javanese dictionaries, a few number of dictionaries is available in the market. It ranges from short-pocket dictionaries to standard dictionaries both in minilingual or in bilingual version. For this study, we reviewed five standard dictionaries that are used by student in our university. These dictionaries are *Kamus Unggah-Ungguh Basa Jawa* (Harjawiyana & Supriya 2009), *Kamus Basa Jawa* (Balai Bahasa Yogyakarta 2003), *Javanese English Dictionary* (Robson &Wibisono 2002), *Bausastra: Kamus Jawa-Indonesia* (Prawiroatmojo 1957), and *Baoesastra Djawa* (Poerwadarminta 1939). The works of Robson & Wibisono (2002), Prawiroatmojo (1957), and Harjawiyana & Supriya (2009) are considered as bilingual dictionaries. However, Harjawiyana & Supriya (2009) is a specialized dictionary focusing on Javanese registers. Poerwadarminta (1939) and Balai Bahasa Yogyakarta (2003) are monolingual dictionaries.

In addition to the use of corpora, most lexicographer now agree that dictionary compilation should consider the users' needs (Lew 2011). For this reasons, before starting the lexicographic work, we conducted a survey study of dictionary use among our students at Javanese Studies Program, Faculty of Humanities, Universitas Indonesia. This research aims to understand how students use Javanese dictionary for their study and how students show

their needs with regard to the microstructure of a dictionary? This information is very important to compile an ideal learner dictionary for our Javanese students in the future.

Method

This research applied a descriptive qualititative approach focusing on student uses and preferences of Javanese dictionaries. For data collection, this research made use of a questionnaire with a Likert scale that is distributed among students of Javanese studies program at Universitas Indonesia. Aside from a list of questions, the questionnaire also provided with a list of images presenting dictionary entry examples to ease students in assessment and evaluation of the selected dictionaries. Only 31 from 40 sets of questionnaire are returned by respondents.

Results

Based on the analysis of 5 dictionaries, namely Javanese English dictionary (Stuart Robson and Singgih Wibisono), *Kamus Basa Jawa*, *Baoesastra Djawa* (Poerwadarminta), *Kamus Unggah-Ungguh Basa Jawa* (Haryana and Supriya), and *Bausastra Jawa-Indonesia* (Prawiroatmojo), each has advantages and disadvantages in the dictionary's microstructure and macrostructure. In detail, the microstructure and macrostructure components of the five dictionaries can be seen in Table 1. In addition, based on the questionnaires distributed to the Javanese Literature students of University of Indonesia, it appears that Javanese Literature students prefer *Kamus Basa Jawa* (Balai Bahasa Yogyakarta) to the other four dictionaries. The results can be seen in Table 2 and Table 3.

Discussion and Conclusion

Dictionary Structure

In this section, the five dictionaries are analyzed based on the microstructure and the macrostructure of the dictionary. However, before the microstructure and macrostructure analysis, the researchers divide the five dictionaries into 2 parts, namely monolingual dictionary and bilingual dictionary. The monolingual dictionary consists of *Kamus Basa Jawa* (Balai Bahasa Yogyakarta) and *Baoesastra Djawa* (Poerwadarminta). Meanwhile, the bilingual dictionary consists of *Javanese English Dictionary* (Struart Robson and Singgih Wibisono), *Kamus Unggah-Ungguh Basa Jawa* (Haryana and Supriyana), and *Bausastra: Kamus Jawa-Indonesia* (Prawiroatmojo). Both types of dictionary (monolingual dictionary and bilingual dictionary) are commonly used by students of Javanese Literature in learning Javanese language. Here is the discussion of the five dictionaries.

a. Microstructure and Macrostructure (Monolingual Dictionary)

In the monolingual dictionary, the definitions on *Kamus Basa Jawa* and *Baoesastra Djawa* are presented in Javanese. According to the age of dictionary, *Kamus Basa Jawa* is younger if compared to *Baoesastra Djawa* which uses old spelling. On the microstructures of the dictionary, these two dictionaries are not that much different. In both dictionaries, every word entered as an entry is labeled with an abbreviation that denotes a variety of uses based on speech levels, such as *ngoko* (n), *krama* (k), and *krama inggil* (k.i). In *Baoesastra Djawa*, an abbreviation that shows the variety of uses is written in lowercase, whereas in *Kamus Basa Jawa* is written with capital letters. The disadvantages of these two dictionaries in terms of microstructure lie in the absence of information on word classes, pronunciation, synonyms, antonyms, and sentence examples.

In the macrostructure section of the dictionary, there is a difference in the entries compiling between *Baoesastra Djawa* and *Kamus Basa Jawa*. At *Baoesastra Djawa*, the entries are arranged only in alphabetical order. Meanwhile, in the *Kamus Basa Jawa*, the entries are not only arranged in alphabetical order, but also arranged based on the

meaning of the entry. Therefore, in the *Kamus Basa Jawa*, the same vocabulary can become two different entries if the definitions are different. In addition, the other difference lies in the subentry arrangement. In *Baoesastra Djawa*, the newly formed word that may become subentry is not placed under the entries, but rather into a single entity with definitions. Therefore, one entry looks full. Meanwhile in *Kamus Basa Jawa*, the subentry is neatly placed under the entry. It eases the readers to find the formed-word. At the beginning of these two dictionaries, the user manual of the dictionary is presented that contains phoneme usage, affix usage, type of words in the dictionary, and various abbreviations used in the dictionary.

Overall, there are similarities and differences in the construction of microstructure and macrostructure of these two dictionaries. Both similarities and differences are almost alike, especially in the construction of microstructure. Both dictionaries are not equipped with word class information, pronunciation, synonyms, and sentence examples on each entry. In addition, the newly formed word that exists on each entry is incomplete. The display example of *Kamus Basa Jawa* can be seen in Picture 1 and *Baoesastra Djawa* in Picture 2.

Picture 1. *Kamus Basa Jawa*

[1]**lungguh***KN* sawah (pelemahan) sing diparingake nggadhuh marang priyayi (abdi dalem) pametune pinangka bayare
[2]**lungguh***N*, **lenggah***K***1** linggih; **2** pepangkatan; pangkat; **3** kaanane (genahe) mungguhing prakara; **4** mapan; wis becik tmr pasang rakiting ukara lsp;
ngglungguhi1 linggih ing; **2** netepi dhawuh; **3** nindakake apa-apa ngetrepi kaya pepanggiling pangkate

Picture 2. *Baoesastra Djawa*

loenggoeh n. **lenggah** k: **1** engg. Linggih; **2** sawah (palemahan) sing pinaringake nggadhoeh marang pijaji (abdi-dalem) pametoene minangka dadi bajare; **3** pepangkatan, pangkat; **4** kaanane (genahe) moenggoehing prakara; **5** mapan, wis betjik tmr. Pasang rakiting oekara lsp; **ng-ꦤ-i**: **1** loenggoeh (linggih ing); **2** netepi dhawoeh; **3** nindakake apa-apa ngetrepi kaja pepanggiling pangkate; ktj. Linggih, piloenggoeh

The examples of dictionary display in Picture 1 and 2 illustrate the compilation of entry *lungguh*. In Picture 1, there are two entries *lungguh* with one subentry *nglungguhi* placed below the second entry of *lungguh*. The two entries are differentiated by definition. Meanwhile, in Picture 2 there is only 1 entry *lungguh* which combine 2 different definitions. In addition, the newly formed word *nglungguhi* is arranged into one, alongside with definition.

b. Microstructure and Macrostructure (Bilingual Dictionary)

In bilingual dictionaries, the definitions are presented in different languages. In the *Javanese English Dictionary*, the entry definition is presented in English. Meanwhile in *Kamus Unggah-Ungguh* and *Bausastra: Kamus Jawa-Indonesia*, the entries' definition are presented in Indonesian. The definitions in all three dictionaries are brief and simple. The presentation of dictionary microstructure of these three dictionaries looks different. In the *Javanese English Dictionary* and *Kamus Unggah-Ungguh*, entries are labeled with abbreviations that indicate the variety of usages based on the level of speech. The abbreviation of *Javanese English Dictionary* are *ngoko* (ng), *krama* (kr), and *krama inggil* (ki), while the abbreviation of *Kamus Unggah-Ungguh* are *ngoko* (n), *ngoko alus* (na), *krama* (k), *krama alus* (ka), and *krama inggil* (ki). Meanwhile, the entry on the *Bausastra*: *Kamus Jawa-Indonesia* is not labeled with these abbreviations. The similarity and disadvantages of these three dictionaries in terms of dictionary microstructure are the

absence of word-class labeling and pronunciation. Of the three dictionaries, only the *Kamus Unggah-Ungguh* that provide an example of the use of words in sentences based on speech levels and also being completed with translations in Indonesian.

In the dictionary macrostructure, the entries of all three dictionaries are arranged in alphabetical order. Only *Kamus Unggah-Ungguh* composes the entries based on the alphabetical order and the meaning of the entries. In the subentry section, the subentry arrangement is placed at the bottom of the entry. Of these three dictionaries, the subentry in *Javanese English Dictionary* and *Kamus Unggah-Ungguh* are more complete. However, both dictionaries have different subentries. The user manual of *Javanese English Dictionary* contains the use of affixes on verbs, adjectives in Javanese language, and spelling in Javanese language. In *Kamus Unggah-Ungguh*, the user manual of dictionary contains an explanation of the entries and subentry, spelling and letters, and abbreviations used in. Meanwhile, *Bausastra: Kamus Jawa-Indonesia*, the user manual only contains the sound descriptions of the Javanese language.

Based on the above explanation, there are disadvantages and advantages in each dictionary. The disadvantages are almost identical to the disadvantages in the monolingual dictionary, i.e, the absence of word-class labels, synonyms, pronunciation, examples, and incomplete subentry. *Javanese English Dictionary* display can be seen in Picture 3, *Kamus Unggah-Ungguh* in Picture 4, and *Bausastra: Kamus Jawa-Indonesia* in Picture 5.

Picture 3. Javanese English Dictionary

Lungguh*ng*, **linggih***ng*, *kr*, **lenggah***k.i.* to sit;
Lunggah-lungguh to keep standing and sitting again;
Nglungguhi to sit on;
Nglungguake to seat s.o.;
Kalungguhan position, situasion, office;
Lungguhan1 seat, place to sit; **2** to be in a seated position ;**3** (*or* lelungguhan) to sit around;
Nglungguhi klasa gumelar *prov* to inherit valuables;
Nglungguhi klasa pengulu *prov* to marry the husband of one's deceased sister;
Palungguhan1 seat, place to sit; **2** position, office.

Picture 4. Kamus Unggah-Ungguh

Lungguh (g) (I) lenggah b: linggih duduk
Ki: pinarak, lenggah (v)
Lungguh, (II)
1) *Kalungguhan* kalenggahan pangkat pangkat, kedudukan
2) *Palungguh, pelungguh (p), lungguh (p)*
palenggah, pelenggah (p), lenggah (p)
sabin utawi pasiten ingkang dipun-gadhuh dening punggawaning panguwaos ing dhusun
sawah atau lahan yang disanggam/ dipinjam oleh pegawai penguasa di desa
3) *Pilungguh* pilenggah ki: dedalemdedalem, dedunung, gegriyabertempat tinggal

3)n: pangebektiku katur Bapak Martana kang pilungguh ing desa Gamping.
K: pangebekti kula katur Bapak Martana ingkang pilenggah ing dhusun Gamping.
I: Hormat baktiku kepada Bapak Martana yang bertempat tinggal di dusun Gamping.

Picture 5. Kamus Bausastra Jawa-Indonesia

Lungguh: duduk; tanah/ sawah jabatan; pangkat, martabat; jawatan
- é prekara : duduknya prekara
- Di – i: dididuduki; diindahkan
- Ng – i: duduk di; pada tempatnya; kena benar

Based on the above analysis, it can be observed the construction of microstructure and macrostructure of the five dictionaries. Although each dictionary has its own advantages, it can generally be concluded that there are still many disadvantages in all five dictionaries. The disadvantages are: the absence of word-class labels, pronunciation, synonyms, usage examples, diverse (dialect), and incomplete formed-word (subentry).

Things to consider in the Javanese language dictionary, either monolingual or bilingual are the speech level and Javanese language dialect. There are three speech levels of Javanese language, namely *ngoko*, *krama*, and *krama inggil*. The use of these three levels of speech differs by context, so the context of usage is required to understand the referred word. Besides, the word formed from each level of speech is also different. Therefore, a complete subentry that contains the word formation is also strongly needed. In the case of Javanese language dialect, the Javanese language dictionary also needs to provide information on the dialect of the word entered as an entry in the dictionary. It will certainly help people who want to learn Javenese language.

Table 1. Microstructure and Macrostructure

No	Dictionary Name	Microstructure							Macrostructure			
		Definition	pronunciation	Word class	Synonym	Register	Example	Typography	alphabetic	Subentry	Compound word	guideline
1	Kamus Basa Jawa	√	-	-	-	√	-	-	√	√	-	√
2	Baoesastra Djawa	√	-	-	-	√	-	-	√	-	-	√
3	Javanese-English Dictionary	√	-	-	-	√	-	-	√	√	√	√
4	Kamus Unggah-Ungguh	√	-	-	-	√	√	-	√	√	-	√
5	Kamus Bausastra Jawa-Indonesia	√	-	-	-	√	-	-	√	√	-	√

Result of Questionnaire

In addition to the analysis of microstructure and macrostructure on all five dictionaries, this study also made use a questionnaire to see students' appraisal of the five dictionaries. The respondents are Javanese Literature students that consist of native speakers and non-native speakers.

There are four components of the assessment of the five dictionaries, namely the function, practicality, completeness, and beauty of the display. Based on the questionnaire, the results are: *Kamus Basa Jawa* has the highest value among the five dictionaries, while *Baoesastra Djawa* has the smallest value. This can be seen in Table 2.

Table 2. Result of Questionnaire (Native and Nonnative Speaker)

No	Dictionary names	Result
1	**Kamus Basa Jawa**	**14,81**
2	Baoesastra Djawa	13,71
3	Javanese English Dictionary	14,19
4	Kamus Unggah ungguh	14,13
5	Kamus Bausastra Jawa-Indonesia	13,90

The comparison between *Kamus Basa Jawa* and *Baoesastra Djawa* in the previous discussion shows that in terms of display, *Kamus Basa Jawa* is more superior to *Baoesastra Djawa*. Besides, *Kamus Basa Jawa* is also younger than *Baoesastra Djawa*. Nevertheless, they both show disadvantages in term of microstructure construction of the dictionary.

Meanwhile, if separated between native speakers and nonnative speakers, based on a questionnaire filled by native speaker students, the results says that *Kamus Basa Jawa* has the highest value, while the *Bausastra: Kamus Jawa-Indonesia* has the lowest value among the five dictionaries. See Table 3.

Table 3. Result of Questionnaire (Native Speaker)

No	Dictionary names	Result
1	**Kamus Basa Jawa**	**14,81**
2	Baoesastra Djawa	13,72
3	Javanese English Dictionary	13,90
4	Kamus Unggah ungguh	13,27
5	Kamus Bausastra Jawa-Indonesia	13

Based on these results, it reflects that for the students of Javanese Literature who is a native speaker, *Bausastra: Jawa-Indonesia* is not very good in terms of function, practicality, completeness, and beauty of the display. Compared to the bilingual dictionary *Javanese English Dictionary* and *Kamus Unggah-Ungguh*, it is clear that *Bausastra: Kamus Jawa-Indonesia* has an incomplete microstructure arrangement.

Conclusion

Based on the analysis of dictionary structure of the five dictionaries of Javanese language, it can be concluded that there are parts of dictionary that need to be considered both in terms of the construction of microstructure and macrostructure. In addition, based on the results of the questionnaire, it can be concluded that the function, practicality, completeness, and beauty of the display become an important thing for the learners of Javanese language. Based the two analysis that have been done, the improvement of Javanese dictionary is required in order to fill the deficiencies/gaps of the previous dictionaries. Therefore, the results of this research can be used as a reference for researchers in compiling and developing a more complete Javanese language dictionary.

References

Amalia, Dora. 2014. Formulasi Pendefinisian dan Model Pengentrian Verba dalam Kamus Pemelajar Bahasa Indonesia (Verb Defining Formulation and Entry Model in Indonesian Learner Dictionary). Dissertation Thesis in Doctorate Program, Department of Linguistics, Faculty of Humanities University of Indonesia.

Atkins, S.B.T. & Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Balai Bahasa Yogyakarta. 2003. *Kamus Basa Jawa*. Yogyakarta: Kanisius.

Chen Yuzhen. 2010. Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries, *International Journal of Lexicography*, 2010, vol. 23 3: pp. 275-306)

Chon Yuah Vicky. 2009. The Electronic Dictionary for Writing: A Solution or a Problem?, *International Journal of Lexicography*, 2009, vol. 22, 1: pp. 23-54.

Dolezal Fredric Thomas, McCreary Don R. 1996. *Pedagogical Lexicography Today: A Critical Bibliography on Learners' Dictionaries with Special Emphasis on Language Learners and Dictionary Users*. Lexicographica Series Maior 96. Tübingen: Niemeyer

Harjawiyana, H. and Supriya, T. 2009. *Kamus Unggah-Ungguh Basa Jawa*. Yogyakarta: Kanisius.

Hulstijn Jan H, Atkins Beryl T Sue. Atkins Beryl T Sue. Empirical Research on Dictionary Use in Foreign-Language Learning: Survey and Discussion in*Using Dictionaries. Studies of Dictionary Use by Language Learners and Translators. Lexicographica Series Maior 88*. Tübingen: Niemeyer, pp. 7-19.

Humblé Philippe. 2001.*Dictionaries and Language Learners*. Frankfurt am Main: Haag und Herchen

Lew Robert. 2004.2004. *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex

Nesi, Hilary. 2013. Researching User and Uses of Dictionaries. In Howerd Jason (ed.). *The Bloomsbury Companion To Lexicography*. London: Bloomsbury Companion.

Poerwadarminta, W.J.S. 1939. *Baoesastra Djawa*. Batavia: J.B. Wolters.

Prawiroatmojo, S. 1957. *Bausastra: Kamus Jawa-Indonesia*. Surabaya: Express & Marfiah.

Robson, S. and Wibisono, S. 2002. *Javanese-English Dictionary*. Hong Kong: Periplus Edition.

Simons, Gary F. and Charles D. Fennig (eds.). (2017). *Ethnologue: Languages of the World, Twentieth Edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Tarp Sven. 2009. Reflections on Lexicographical User Research, *Lexikos*, 2009, vol. 19: pp. 275-296

Welker Herbert Andreas. 2010. Dictionary Use: A General Survey of Empirical Studies. Brasilia: Author's Edition.

A discourse analysis of editors' prefaces of (Chinese &English) bilingual dictionaries

Wai-on Law
The Chinese University of Hong Kong
wolaw@arts.cuhk.edu.hk

Abstract

The preface is a kind of 'paratext' (Genette, 1997). As a mediator between the dictionary text and the user, the editor can use it to recount the editing process, the publication reasons, and the book features. Probably the least read of the dictionary, both the preface and the user's guide have been off the users' radar (c.f. Law, 2010). Yet, a discourse analysis of this academic sub-genre can generate insights into the editing strategies, the historical background, and ideological influence on the editing process. More substantial findings about this text type are deducted, e.g., its functions, themes, format, writing style, and the editor's role. This study employs the model of Bhatia (1993) and the "move" notion for academic article introductions by Swale (1990) on 60 editors' prefaces from 47 bilingual English-Chinese and Chinese-English dictionaries from the past five decades in three Chinese communities. The software AntConc is also used to help shed light on the keyword frequency, themes, and other language characteristics. From the data, 9 discourse moves are summarized, and the following main functions concluded: interpretive, instructive, informative, affective and vocative. Linguistic statistics evidence the straight-forward and informative characters of these functions. Diachronically, the themes of these prefaces have become more comprehensive, and their editing process more prominent. It is suggested that more emphasis be placed on this text type to show the editing aims, principles, process and publication background, which could contribute to the tracking of lexicography development. Innovation in its format could enhance users' effective use of the main text.

Keywords: bilingual dictionaries, editor's preface, paratext

1. Research background

Genre is an established communication means, commonly used in professional or academic communities, with definite and recognized communication purposes, and strict structure and norms (Swales, 1990; Bhatia, 1993). The genre that professional or academic bodies use is in close connection with their methodologies, and the manner of presenting their messages are aligned with their norms, beliefs and ideologies (Berkenkotter & Huckin, 1995; quoted in Bhatia 1997). The communication purpose defines the generic structure; a different purpose leads to the development of anther genre or a sub-genre. The structural and pragmatic differences lie in the genre-specific lexis, grammar and rhetorical devices (Bhatia, 1993). Discourse analysis studies the language expression in professional or academic settings, with four emphases: (1) pragmatic knowledge; (2) generic knowledge; (3) structural identification; and (4) genre mastery (Bhatia, 1997). Bhatia categorized academic introductions as an academic genre aiming at introducing an academic work. The editor's preface of a dictionary belongs in the introduction genre as a sub-genre.

French literary critic G. Genette (1997) was the first person to propose the concepts and an analysis of the paratext in a systematic manner. According to him, the purpose of a paratext is to make the reader more receptive to the main text, and to guide them to a proper reading. Examples include author's preface, afterword, headings, notes, illustrations, author's interviews, personal correspondence. As a sub-genre of the paratext, the preface points out the significance of the theme of the main text, and the value of the content. The characteristics of the paratext are dependent on its location, temporal background, form, communicative means, writer and audience, as well as functions.

Cai and Pei (2015) opined that Genette's study was synchronic instead of diachronic in terms of the form, content and functions of paratexts. Genette's methodology was inductive when categorizing paratexts, but not comprehensive and systematic enough. Geng (2016) commented that researches on paratexts were primarily qualitative in methodology, and case studies were done from a cultural or sociological perspective, descriptive of the format of the paratext in point, and examining its role, functions and significance under a specific socio-cultural system. A small number of researches began to work on the linguistic features of paratext types like the foreword and the afterword. Large-scale quantitative studies are still rare.

The preface or foreword of the dictionary has long been neglected, as both the reader and the researcher just focus on the main body of the lexicographic work. It is rarely read or researched into as a paratext. Lexicographic history is a regular research topic, but editors of dictionaries seem almost invisible. Yet, the direct expression of the editor clearly indicates the editing purpose of the dictionary, the principles, process, and the publishing background, all rich material to the history of lexicography. According to Bhatia (2006), the preface of the dictionary can be classified as an academic paratext, with its special format, functions and language features. These can be validated in the examples in this study.

2. Methodology

Based on Bhatia's (1993) genre analysis framework, this study approaches the genre style and features of the editor's preface of dictionaries, including the organisation, theme, and content. He put forward five steps for analyzing an unfamiliar genre. (1) Survey the existing literature. (2) Investigate the text setting of that genre, including the author, readership, their relationship and goals; the different background of the community where the textual discourse is situated; the related language conventions of the text. (3) Select genre or sub-genre language material for analysis based on the communication purpose, general setting and features. (4) Examine the pragmatic setting of that genre, including various pragmatic

norms. (5) Analyse linguistic features: a. lexical units, grammar; b. syntactic; c. structure, in reference to the analytical model for academic article introductions by Swales (1990).

Besides, new sets of statistical data are categorized by the software application AntConc, e.g., frequencies of keywords. The figures can help define the features of this paratext type, and explore the themes and focuses of editors of different historical background. The existing corpora contain examples of a vast variety of genres, and are thus incomparable with this one regarding keyword analysis.

From 1841 to 2004, 109 English/Chinese bilingual dictionaries were published in Hong Kong (Chan, 2005). In China, the number of bilingual dictionaries produced in the recent five decades reached 1,800, among which 70% were English/Chinese and Chinese/English (Zhang and Huang, 2000; quoting from Li, 2003). The samples of this study were taken from the editors' prefaces of the English/Chinese and Chinese/English bilingual dictionaries in the last 50 years.

In total, 105 dictionaries in the recent half century were sampled, among which 64 were English/Chinese in language direction, and 41 from Chinese/English. For comparability, only comprehensive bilingual language dictionaries were selected, but not small-scale (mini) ones or glossaries. The origins of publication were Mainland China, Hong Kong and Taiwan. The basic figures of the samples are listed below.

1. Bilingual dictionary samples*	47 (100%)
2. Publication period	
a. 1960s	5 (10.6%)
b. 1970s	6 (12.8%)
c. 1980s	9 (19.1%)
d. 1990s	8 (17%)
e. 2000 till now	19 (40.4%)
3. Language direction: English/Chinese or Chinese/English	
a. English/Chinese	24 (51.1%)
b. Chinese/English	20 (42.6%)
c. English/Chinese and Chinese/English	3 (6.3%)
4. Editor's preface samples**	60 pieces
5. Average word count of each preface	2027
6. Sample with the smallest word count in the preface: *A new complete Chinese-English dictionary* (1966)	298 words
7. Sample with the greatest word count in the preface: *The Chinese-English dictionary* (2015)	8044 words

Table 1: Statistic summary of English/Chinese and Chinese/English bilingual dictionary samples

*Among the 105 examined dictionaries, 50 of them contain (an) editor's preface(s); the prefaces of 3 dictionaries are duplications of other 2 dictionaries, and were not included.
**Only including prefaces or forewords from the editors, publishers or revisers; excluding those from other people not listed as editors, or user guides; several items have bilingual prefaces, and only one of them would be selected. As well, a few English/Chinese dictionaries from 1999 to 2000s were translated from an original version (i.e. bilingualized), and thus only the source English prefaces were chosen. 13 items contain 2 or more prefaces.

The prefaces come under different nomenclature in Chinese. They are collectedly called 'the preface' or the 'editor's preface' in the study.

3. Results and discussion

3.1 Thematic analysis

In reference to the Preface to Third Edition of the *Oxford English Dictionary* (Simpson, 2000), and Swales's (1990) framework on the introductions of research articles, 5 moves were originally drafted. But after thorough reading of the 60 pieces, the moves and themes of the 47 sample prefaces from 47 dictionaries are revised and summarized as below.

Move	Number of dictionaries with this move
1. Publication or reader needs	38 (80.9%)
2. Reference of related dictionaries	33 (70.2%)
3. Editing process, history	32 (68%)
4. Editing principles, editor's reflections	25 (53.2%)
5. Introduction to the arrangement of the dictionary, its features	44 (93.6%)
6. Mention of the characteristics of varieties of languages (e.g., American English, Australian English, the pronunciation systems of Chinese and English)	16 (34%)
7. List of editorial team, acknowledgments	32 (68%)
8. Limitations of the dictionary	31 (66%)
9. Invitation to the reader for corrections	35 (74.5%)
10. Other (ideology)	1 (2.1%)

Table 2: Numbers of moves in dictionary prefaces

Examples of moves are quoted below. The quotations are translated from Chinese by the researcher, unless specified.

(1) Publication or reader needs: 'What the cultural circle needs today is a comprehensive Chinese to English dictionary. The market does offer a variety of choices, yet most content is outdated and simplistic…. the source material is definitively unsuitable, while the indexing is far from handy.' (Lee, S.T. (Ed.) (1966). *A new complete Chinese-English dictionary.* Hong Kong: China Publishers Co.)

(2) Reference of related dictionaries: 'The *Kongxi dictionary* reprinted under the imperial order during the Daoguang regime [1820-1850] was the latest published dictionary of the Commercial Press.' (Li, Y. (Ed.) (1966). *A new Chinese-English dictionary*. Taipei: Commercial Press.)

(3) Editing process, history: 'To edit a Chinese dictionary like the *Concise Oxford dictionary* has been my decades-long dream. In 1944…. 1966… after seven years of dedicated efforts, the dream final came true.' (Lin, Y. (Ed.) (1972). *Lin Yutang Chinese-English dictionary of modern usage*. Hong Kong: The Chinese University of Hong Kong.)

(4) Editing principles, editor's reflections: 'The entries of *Hanying cidian* (new century edition) are Chinese-based, with the special expression of English in consideration, so that they demonstrate the inherent rationale of the Chinese to English dictionary.' (Wu, G. (Ed.) (2001). *Hanying cidian* (new century ed.). Shanghai: Shanghai Jiaotong University Press.); 'Besides being a purposeful reader with scholarly talons laid on every available book, what other qualities are required of a lexicographer?' (original quotation; Lu, G. (Ed.) (2015). *The English-Chinese dictionary*. (2nd ed.). Shanghai: Shanghai Translation Publishing House.)

(5) Introduction to the arrangement of the dictionary, its features: 'In our compiling of this multi-functional English to Chinese dictionary, we made an attempt to include pronunciation

marking, definition, usage and etymology.' (Chang, Q., & Tsai, W. (Eds.) (1966). *New English-Chinese dictionary.* Hong Kong: Commercial Press.)
(6) Mention of the characteristics of varieties of languages (e.g., American English, Australian English, the pronunciation systems of Chinese and English): 'In light of the increasing exchange between mainland China, Hong Kong and Taiwan, … an adequate amount of Taiwanese terms are added.' (*Oxford intermediate learner's English-Chinese dictionary* (new 3rd ed.). (2002). Oxford University Press (China).); 'In the modern days, a variety of English has been developed…. That's why *Longman dictionary of English language & culture* covers both the British and American English.' (*Longman dictionary of English language & culture (English-Chinese)* (bilingual ed.). (2003). Hong Kong: Pearson Education Asia Limited.)
(7) List of editorial team, acknowledgments: 'The publication of this book was made possible due to the voluntary contributions of many individuals and communities…. I hereby dedicate my heartfelt gratitude to all those mentioned.' (*ABC Chinese-English comprehensive dictionary*, J. DeFrancis (Ed.) (2003). Shanghai: Hanyu dacidian chubanshe.)
(8) Limitations of the dictionary; (9) Invitation to the reader for corrections: 'Due to time constraints, there may be some areas that are not as good as we desire. We hope readers could give us their precious comments.' (T. Han, & Y. Li (Eds.) (2003). *Jingbian Yinghan Hanying Cidian*. Beijing: Zhongguo dabaikequanshu chubanshe.)
(10) Other (ideology): 'In the process of its preparation, we studied time and again our great leader Chairman Mao's teachings…, and unfolded a revolutionary mass criticism of the philosophy of servility to things foreign, scholasticism, and the scum of feudalisim, capitalism and revisionism which abounds in old-type English-Chinese dictionaries….' (original quotation; Editing Group (Ed.) (1975). *A new English-Chinese dictionary.* Hong Kong: Joint Publishing Company.)

The moves above carry the following functions: (1) interpretive, to enable readers to understand the features, publishing purpose and arrangement of the dictionary; (2) instructional, to point out the editing principles; (3) informative, to cite related dictionaries, editing team, language features of English and Chinese, the editing process and history; (4) affective, to acknowledge all those who have partaken and assisted in the publication; (5) vocative, to invite readers for direct comments and corrections. The number of functions is greater than that of the thematic functions under Dimitriu's (2009) study of translators' prefaces, probably due to the tremendous time and manpower involved in the publication. The language features of dictionary prefaces tend to be straightforward and rational, slightly formal, with linguistic terms, in reference of related dictionary works. These qualities are in line with the academic genre described by Bhatia (1997). As preface to a reference tool, such sub-genre is concentrated on concrete content of the work instead of theories. Professionally, since language is involved comprehensively in daily life, most editors admit their limitations in knowledge, and encourage two-way communication. Besides, as a dictionary is mostly out of co-operative efforts, team work is often acknowledged. In these various aspects, they differ from academic articles.

3.2. Linguistic analysis

The software application AntConc was in use to count word frequencies. Regarding MacDonald's (2002) suggestions, on the syntactic level, the majority of the 60 pieces employ concrete nouns, but few abstract ones. The top 10 nouns on the list are (again, translated from the source Chinese): 'dictionary' (847 times), 'English' (478), 'readers' (237), 'lexicon' (228), 'language' (187), 'Chinese/English' (154), 'China' (136), 'word' (125), 'English' (124), 'edition' (112). The 3 most seen abstract nouns are: 'necessity' (107), 'culture' (106),

and 'meaning' (62). The frequency of abstract nouns is much lower than that of concrete ones. This propensity to concrete nouns is closely related to the themes.

For pronouns, 'we' appears 261 times in total, while 'I' only 130 times. Consider that Chinese sentences can start without a subject, and the editor-writers could have taken the dictionary as the subject instead, and so avoiding the 'we' and 'I' , the use of the first-person plural pronoun is even more significant. It reflects the tedious undertaking as group work, and thus the preface speaks for the whole team, and that lexicographic knowledge is commonly shared. Three sample dictionaries were named after individuals, e.g., *Lin Yutang Chinese-English dictionary of modern usage* (1972), *Liu's Chinese-English dictionary* (1978), *An advanced English dictionary* (2005) (named after 'Zhang Daozhen' in the Chinese title), and one singled out a single person as the editor, Tan Chor Eng, for *A draft copy of modern Chinese-English dictionary* (1973). Their prefaces do use 'I' more. Alternatively, the term 'the Editor' is also used 90 times, which does not specify the singular or plural in Chinese grammar.

The content in language dictionaries is closely hinged on language development. This is reflected in the frequent use of temporal lexis in prefaces, e.g., 'times' (70 times), 'contemporary' (51), 'modern' (49), 'century' (45), evident of the synchronic significance. Quotations are seldom made. The few exceptions are: 'Zhu Chunsheng was insightful in his *Shuowen tongxun dingsheng* [a philological classic in the Qing dynasty]' (*Lin Yutan Chinese-English dictionary of modern usage*, 1972); 'I begin by quoting Samuel Johnson's remark when he commented on the widespread rural illiteracy in Scotland in his time….' (original quotation; Lu, G. (Ed.) (2015). *The English-Chinese dictionary*. (2nd ed.). Shanghai: Shanghai Translation Publishing House.) A dictionary is both a reference tool and a commodity, a well of linguistic data and practical knowledge, with few personal opinions; yet citing other lexicographers on editing dictionaries can share the subject knowledge. Another relevant word use data set is that, 'readers' appears 237 times, 'learners' 53 times, and 'users' 26. The purpose for editing reference tools is to provide language information for readers' regular use, and thus their needs and this relationship are emphasized.

On the meta-discourse level, some findings were made according to Kopple's (2002) language strategies. The mostly seen textual connectives are: 'and' (*he*; Hanyu Pinyin) (573 times), 'and' (*yu*, the more classical form) (205), 'or' (*huo*) (164), 'and' (*ji*, a synonym) (146), 'but' (*dan*) (117), 'besides' (*ping*) (109), 'also' (*hai*) (80), 'as well as' (*yiji*) (79), 'because' (*yinwei*), 'because of that' (*yinci*) (63). Most of these connectives link up parallel structure; only the last two are causal. This characteristic shows that prefaces are primarily informative in nature, and secondarily expository, with other relationships like contradiction, supposition, concession, used much less often. Comment markers are written for the writer to express intentions and feelings directly, or to speak to the readers. The most frequent wording is 'hope' (27), 'satisfied' (27), 'obliged' (19), in sync with move 9.

3.3 Diachronic analysis

Thematic evolvement in dictionary prefaces is observed despite the small sample size. The statistics are presented as follows.

Publication decade	Number of dictionaries	Most frequent first three moves (by frequency)*	Average move number**
1960	5	5, 1, 9	4.8
1970	6	1, 3, 2, 4, 5, 8, 9 (the last 5 ranked the same)	5.2
1980	9	5, 7, 8	6.3
1990	8	1, 5, 2, 7, 8, 9 (the last 4 ranked the same)	6.3
2000 on	19	5, 3, 1, 2 (the last 2 ranked the same)	6.5

Table 3: Frequency of move in dictionary prefaces by publication decade

* C.f. Table 2 for the move descriptions.
**Move 10 is 'other', an indefinite item. The base number is thus set at 9.

The figures above show that the moves in dictionary prefaces have become more comprehensive. The first major theme remains to be introduction of the arrangement and features of the dictionary, the second being the purpose of publication, and the third acknowledgment of the limitations of the work, and an invitation for readers' feedback. One noteworthy finding is that the editing process and history has been a prominent feature since 2000. Functionally, it means that the editor's preface is aimed at the following three functions in descending order of importance: (1) interpretive; (2) informative; and (3) vocative.

Out of 47 sample dictionaries, only one (*A new English-Chinese dictionary* cited in Section 3) mentions obvious ideological statement. Actually, *Yingzi rumen* (*An introduction to the English lexicon*) (1874) also contained an example, '… Life in foreign settlements is decadent and extravagant…. Once in power, people are indulged in gambling and in brothels regardless of the costs…. My humble wish is that readers could learn from these lessons….' Such comments were normally not made when the society was in stability, when dictionaries focused on the reader's individual needs.

A comparison with the 84 translators' prefaces of contemporary English literary works investigated by McRae (2012) demonstrates many common themes, e.g, the cultural and historical background of the text, acknowledgments, the limitations of translators, the translator's role as editor, some grammatical rules, the disparities between the American and British English, and the reader's responsibility.

4. Conclusion

This study has examined 60 editors' prefaces from 47 English/Chinese and Chinese/English bilingual dictionaries published in the recent 5 decades. From a discourse analysis of the themes are drawn 9 moves for 5 major functions: interpretive, instructional, informative, affective and vocative. Their plain and informative style befits these functions. Diachronically, the themes have become more comprehensive, with greater emphasis on the editing process. It is suggested that publishers give due credit and emphasis to the editor's preface given its direct content about the editing purpose, principles, process and publication background, which could contribute to the history of lexicography and pedagogical lexicography, so that readers can better understand the reference tool they use.

Apart from its contributions to lexicography, the study is also pertinent to discourse analysis, to which a set of data of a paratext type is provided, with qualitative and quantitative methods. The framework proposed by Bhatia (1993) and the moves for introductions of

academic articles by Swale (1990) were applied with some modifications for such academic reference as the dictionary, in addition to linguistic statistics and diachronic and synchronic considerations. This methodology could be applied to similar studies.

Undeniably, the dictionary preface and its writer have long been neglected by users. When most dictionary users resort to the internet or their mobile phones, directly tapping their search on their devices, the editor's preface seems left for the experts to read, no matter how well-written. Given this, perhaps the dictionary editor and the publisher could spend more time on the preface web page for more attraction to readers. An interactive online demonstration of all the possible uses of the dictionary could be a good start.

5. Acknowledgment

This project received support from the Arts Faculty of the Chinese University of Hong Kong (Project Number: 505004).

6. References

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Essex: Longman.

Bhatia, V. K. (1997). The power and politics of genre. *World Englishes* (3), 359-371.

Bhatia, V. K. (2006). Analysing genre: Some conceptual issues. In M. Hewings (Ed.), *Academic writing in context: Implications and applications* (pp. 79-92). London: Continuum.

Chan, S. W. (2005). Lexicography in Hong Kong: 1841 – 2004. *The Hong Kong linguist* (25), 8 – 20.

Dimitriu, R. (2009). Translators' prefaces as documentary sources for translation studies. *Perspectives: Studies in translatology*(3), 193-206.

Genette, G. (1997). Paratexts: *Thresholds of interpretation* (Tr. by J.E. Lewin). Cambridge: Cambridge University Press.

Kopple, W. J. V. (2002). Metadiscourse, discourse, and issues in composition and rhetoric. In E. Barton & G. Stygall (Eds.), *Discourse studies in composition* (pp. 91-113). Cresskill, NJ: Hampton Press.

Law, W.-O. (2010). *Translation students' use of dictionaries: A Hong Kong case study for Chinese to English translation*. Doctoral dissertation, University of Durham, *Asian EFL journal*. [Access via: http://www.asian-efl-journal.com/Thesis/Thesis-Wai-on-Law.pdf].

Li, D. F. (2003). Compilation of English-Chinese dictionaries: The user's perspective. *Journal of translation studies*(8), December, 91–115.

MacDonald, S. P. (2002). The analysis of academic discourse(s). In E. Barton & G. Stygall (Eds.).

Discourse studies in composition (pp. 115-129). Cresskill, NJ: Hampton Press.

McRae, E. (2010). The role of translators' prefaces to contemporary literary translations into English: An empirical study. In A. Gil-Bardaji, P. Orero & S. Rovira-Esteva (Eds.), *Translation peripheries: Paratextual elements in translation* (pp. 63-82). Bern: Peter Lang.

Simpson, J. (2000). Preface to the Third Edition of the OED. *Oxford English dictionary* (3rd ed.).

Oxford: Oxford University Press.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Using Dictionaries in Metaphor Identification

Wu Jihong
Guangdong University of Foreign Studies
wujihongs1997@163.com

Abstract

Metaphor has been the focus of cognitive linguistics, psycholinguistics, applied linguistics, corpus linguistics, and metaphor identification lays a solid foundation for metaphor research. Since Lakoff and Johnson (1980) proposed the Conceptual Metaphor Theory, much attention has been given to the conceptual and cognitive dimensions of metaphor, leaving linguistic dimension secondary. However, when MIP was introduced in 2007, which aims to identify metaphorically used lexical units in natural discourses, metaphor researchers have developed a systematic and reliable methodology for identifying linguistic metaphor instead of working with institution and subjective criteria, which enables them to focus their research on different levels-linguistic forms, conceptual structure and cognitive processing. As MIP requires metaphor analysts to work through 4 steps, in which they depend heavily on dictionaries to determine lexical units and specify the basic and contextual senses, the use of dictionaries becomes the critical element in MIP. The Pagglejaz Group chose *Macmillan English Dictionary for Advanced Learners* as reference tool, while MIPVU, the elaborated version of MIP, used *Longman Dictionary of Contemporary English* and *Oxford English Dictionary* apart from MED. The author, by demonstrating the use of different types of dictionaries in MIP, tries to show that together with learners' dictionaries, historical dictionaries, collocation dictionaries and specialized dictionaries can also be used for cross reference to guarantee the reliability of linguistic metaphor identification in MIP.

Key words: metaphor identification; dictionary; linguistic metaphor; sense

Introduction

Since Lakoff and Johnson published *Metaphors We Live By* in 1980, metaphor research has become the focus of cognitive linguistics, psycholinguistics, applied linguistics, and corpus linguistics. With the application of corpora, people begin to emphasis the difference between grammar and specific usage of a language in their research (Steen, 2007), and accordingly, large corpora are used to facilitate metaphor research related to specific contexts, in which metaphor identification becomes a pressing issue (Krennmayr 2013). Metaphor research can be approached from two perspectives: linguistic metaphor and conceptual metaphor, and since MIP was introduced in 2007(metaphor identification procedure, Pragglejaz Group, 2007), the identification of linguistic metaphors has attracted more attention than ever before. In MIP, the most crucial part is the contrast between the basic meaning and contextual meaning of lexical units, and since "a meaning can not be more basic if it is not included in a contemporary users' dictionary" (Steen et al., 2010:35) and 99% metaphorical usages from native speakers can be found in dictionaries of contemporary English (Steen 2011), using dictionaries, usually learners' dictionaries, becomes the key factor in metaphor identification based on MIP. The author, by demonstrating the use of dictionaries in MIP, tries to show that that not only learners' dictionaries, but also historical dictionaries, collocation dictionaries and specialized will help researchers make relatively consistent and objective judgment in linguistic metaphor identification.

1. MIP: a bottom-up approach in metaphor identification

Early in the metaphor study, metaphor identification usually relied on the researchers' intuition, such as Lakoff and Johnson's research (1980). Later on, the development of corpus linguistics enabled researchers to get rid of the dependence and search for a relatively unified and effective standard in metaphor identification. Presently, the metaphor identification mainly includes two kinds of approaches: top-down and bottom-up approaches. The former presets conceptual metaphors, then retrieves corresponding linguistic metaphors from the text, while in the latter no conceptual metaphors are presumed and researchers try to derive mappings from linguistics expressions which they identify as metaphorically used. The identification of metaphor based on the application of the basic principles and methods of corpus linguistics in essence can be categorized as the top-down approach, but in recent years, people come to realize the limitation of this deductive research method: there are no standard procedures to identify conceptual metaphors, and researchers have to rely on their intuitions to a great extent. At present, more and more researchers prefer the bottom-up approach, and MIP and its upgrade MIPVU (Metaphor Identification Procedure at VU University level, Steen et al., 2010) are employed as a typical bottom-up approach (for convenience both are referred to as MIP). In MIP a language unit can be divided into metaphorical and non-metaphorical expressions, and once the semantic consistency is destroyed by introducing the conceptual meaning of a different domain, the language unit can identified as a metaphorical expression.

MIPVU, a revised version of MIP, has made a further improvement in metaphor identification. It extends metaphors to similes and implicit metaphors so there are three types of metaphors in MIPVU: indirect metaphors, direct the metaphors and metaphor indicators, eg, in the sentence *the marriage is a trap,* "trap" is an indirect metaphor; *He eats like a pig*, "pig" an direct metaphor, while *like, as, compare*, etc. are metaphor indicators. Moreover, the lexical unit in MIPVU is refined to its part of speech rather than lemma in MIP. In addition to the *Macmillan English Dictionary for Advanced Learners* (henceforth MED), the reference in MIP, MIPVU also refers to *Longman Dictionary of Contemporary English* (henceforth LDOCE) and *Oxford English Dictionary* (henceforth OED) for help. Perhaps the biggest difference between MIPVU and MIP lies in the fact that in MIPVU it's not enough to make

contrast between the basic meaning and context meaning to identify metaphors, but the semantic references of the two concepts have to demonstrate similarity in the external or function. To a certain degree, MIPVU provides more comprehensive, objective criteria in metaphor identification than MIP.

The specific steps in MIP are as follows:

1) read the entire text–discourse to establish a general understanding of the meaning.
2) determine the lexical units in the text–discourse
3) (a) for each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.
   (b) for each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be
   —more concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];
   —related to bodily action;
   —more precise (as opposed to vague);
   —historically older;
   (c) if the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4) if yes, mark the lexical unit as metaphorical

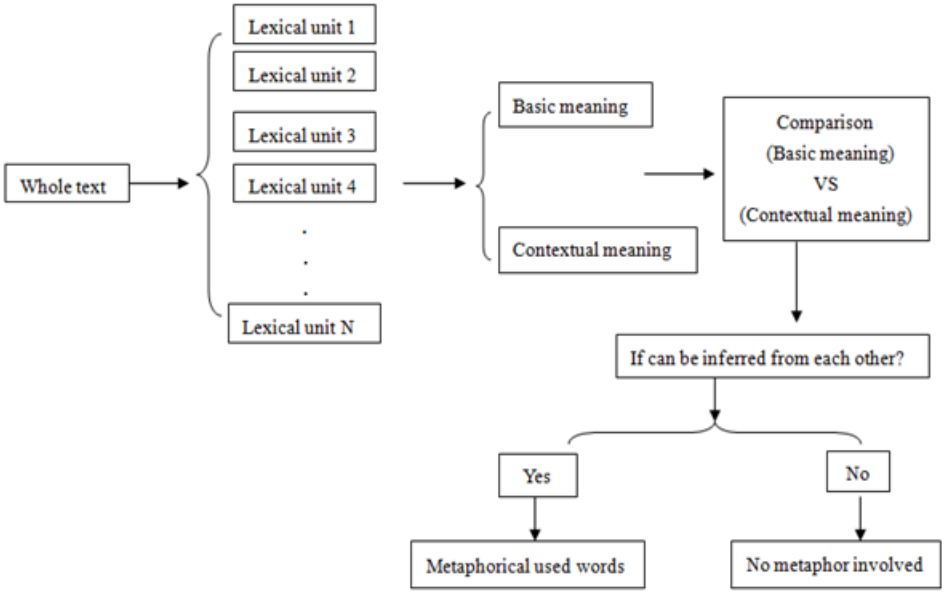


Figure 1 Process of identifying metaphorically used words (Pragglejaz 2007)

As is shown, MIP procedure can only identify linguistic metaphors, restraining from presuming conceptual metaphors. Unlike top-bottom approach, the five-step method of MIP, which restricts itself to dealing with comparing and contrasting meaning as defined in the dictionaries(Steen 2007), helps researchers to get rid of dependence on their intuitions with comparatively reliable basic meaning and contextual meaning of the lexical unit coded in dictionaries. Moreover, as the comparison between the basic meaning and contextual meaning strictly follows dictionary definitions to determine metaphorically used words, providing the basis from which cross-domain mappings are constructed, MIP, with its focus on linguistic

metaphors, prevents researchers from "seeing concrete manifestations of conceptual metaphors everywhere"(Steen 2007:27).

2. Using dictionaries in MIP

MIP strictly adheres to standard English dictionaries to determine the lexical unit and compare and contrast its basic meaning and contextual meaning, so the importance of dictionaries can never be overestimated. As MIP is targeted on formal, contemporary, standard British English (Steen 2007), dictionaries based on a large, general and contemporary English corpus are preferred, mainly learners' dictionaries, though not restrained to them.

2.1 Use of learners' dictionaries in MIP

Pragglejaz Group chose learners' dictionaries for the following reasons: First of all, most contemporary English learners' dictionaries, with no exception, are compiled based on large corpora of contemporary English from different discourses: MED was compiled based on a systematic processed corpus of 220 million words, large enough to provide a number of citations for all but the rarest words, and LDOCE, the Longman Corpus Network, a 330 million word database. Hence they are considered adequate for general language analysis and can fully satisfy the need for metaphor research (Pragglejaz Group 2007). Secondly, unlike dictionaries complied for native speakers, in learners' dictionaries special consideration is given to high-frequency words with exquisite sense divisions, precise definitions, typical examples and collocations. Words like *say, see, light and grasp* etc. create few difficulties for natives but for non-native speakers, and it is high-frequency words rather than difficult or rare words that pose serious problems when they try to differentiate the literal meaning and metaphorical meaning, hence the dictionaries are heavily used in MIP, especially in step 3.

1) definition

Sometimes not only sense division, but also definitions will help distinguish the basic meaning and contextual meaning of a lexical unit. In the following example, the first sense of "embrace" is related to body action, which is concrete, and the second sense, with abstract collocates "idea", "belief" and "opinion" etc. will make it a direct metaphorical sense.

[1] *Community standards may **embrace** moral principles or they may not.*Φ

embrace

MED: to completely accept something such as a new belief, idea, or a way of life [sense 2a]

LDOCE: to eagerly accept a new idea, opinion, religion etc. [sense 2]

2) collocation information

Most contemporary learners' dictionaries allocate considerable space to collocation information as it presents the way a word is used in specific context and with its collocates, we can decide on its meaning, even when the definition is not sufficient to make a judgment.

[2] *He turned round and directed a **torrent** of abuse at me*.

The word "torrent" in MED has two meanings: The first, related to water flow, can be taken as the basic meaning, and the second referring to "a large amount of something, especially something unpleasant" may or may not be deemed to be abstract as "something" is ambiguous though the word "unpleasant" may, to a certain extent, indicate its abstractness. However, if we turn to the highlighted collocation pattern "of a torrent of abuse/words/criticism" in MED for its second sense, we can be fairly assured that

"something" is abstract and "torrent" in the second sense most probably relates to contextual meaning.

2.2 Use of historical dictionaries

The core issue in using MIP to identify metaphor is and above all whether the two senses are listed as two separate, numbered sense descriptions in the dictionary. Though it's believed "the overwhelming majority of cases can be solved by using the Macmillan dictionary, and the Longman dictionary as a second opinion when it is needed"(Krennmayr 2008:107), it is not rare at all that information provided in learners' dictionaries is insufficient for researchers to determine the basic meaning and contextual meaning. For pedagogical purposes, in learners' dictionaries senses are sometimes collapsed and subtle meanings are ignored (Steen 2007; Deignan 2005). And to make things worse, for the target readers, the most frequently used sense of a word would appear as the first sense when its historical development is usually disregarded, which will attribute to the disagreement among researchers concerning the basic sense. Should it occur, information provided in learners' dictionaries will not be sufficient for researchers to make an objective judgment, especially when two meanings are subsumed into one sense description or one of the senses is missing in the dictionaries.

1) sense conflation

Here are some examples:

[3] *It would **use** new methods to teach traditional academic subjects and equip young people with technical skills.*

Our intuition tells us that "use" in "use a method" is different from the one in "use a tool". However, if we consult learners' dictionaries, we will find:

use v.

MED: to do something using a machine, tool, skill and method etc in order to do a job or to achieve a result [sense 1]

LDOCE: if you use a particular tool, method, and the service, ability etc, you do something with that tool, by means of that method etc, for a particular purpose [sense 1]

We will fail to make a distinction between the basic sense and contextual sense as the literal sense and abstract sense are conflated in the exampled sentence, so if we adhere to MIP the word is not metaphorically used, which is against our intuition. Nevertheless, if we turn to OED, a historical dictionary, we will see:

use v.

OED: II. to put to practical or effective use; to make use of, employ, esp. habitually. From the 20th cent. some senses in Branches I. and III. (e.g. senses 3c, 6, and 16) have increasingly been understood ***instrumentally*** as implying particular ends or purposes, even when there is no explicit context of that kind; as a result these uses have converged on the senses in this branch (highlighted by the author)

a. to put (an instrument, implement, etc.) to practical use; *esp.* to make use of (a device designed for the purpose) in accomplishing a task. [sense 8a]

c. to make use or take advantage of (a quality, condition, idea, or other immaterial thing) as a means of accomplishing or achieving something. †Formerly also *intr.* with *of*, (occas.) *with*. [sense 10]

OED makes a segment between "use" related to material things and immaterial things, but citations in OED show that the two different usages occurred nearly at the same time in middle English(about c1300,c=circa) so the concrete usage is not historically older, neither can the semantic relationship be found between the two. If we apply the criterion of MIP, "use" in the example [3] is not metaphorically used.

2) sense omission

As mentioned in previous part, two separate sense descriptions for a lexical unit are considered as a precondition for contrast between the basic meaning and contextual meaning, however, due to the restricted space, it's quite possible that there will be only one sense, usually the most frequently used one listed in learners' dictionaries, while actually there are more than one. Let's see "fervent" and "ardent" in the example [4] and [5]:

[4] *There were **fervent** arguments both for and against gun control.*
[5] *Even his most **ardent** supporters disagreed with this move*.

In MED and LDOCE, both "fervent" and "ardent" have only one sense, which describes emotion, but in OED, besides the one associated with emotion, they both have meanings referring to temperature, with which we can feel confident about their metaphorical usage in the example [4] and [5]:

fervent: hot, burning, glowing, boiling[sense 1]
ardent: burning, on fire, red-hot; fiery, hot, parching[sense 1]

Actually, apart from sense division, the etymological information provided in OED also helps us make judgment: It shows that both "fervent" and "ardent" have Latin origins when Latin "fervent" meant "boil", "glow" and "ardere" meant " to burn", which supports their metaphoricity.

As most learners' dictionaries are based on descriptivism and draw data from corpus, and little consideration is given to etymological information (though in CD-ROM etymology may be provided). On the other hand, in learners' dictionaries frequency is taken as priority in sense arrangement, and people tend to accept the most frequent sense, usually the first one as the basic one, even though it is not necessarily related to its basic meaning (Pragglejaz Group, 2007). To avoid the misjudgment, a historical dictionary becomes a valuable resource in MIP, especially in finding the basic sense:

[6] *I'll just leave the engine running while I go in.*

The highest frequency usage of "leave" is "to go away from a place or a person"(LDOCE sense 1), and most probably, it may be taken as the basic sense. But OED tells us that "leave" was originated from the Old English "bequeath", meaning "allow to remain and leave in place", and still earlier, from German "bleiben", meaning "remain", so, the basic meaning should be "to let something remain in a particular state, position, or condition" (LDOCE sense 5) rather than its first sense, and when we compare and contrast the basic meaning and contextual meaning in example [6], the conclusion can be drawn that "leave" is not metaphorically used.

Moreover etymological information is especially useful for determining the basic meaning of culture-loaded words:

[7] *The students' rooms are **spartan** but clean, with no carpets or central heating.*

spartan
MED: very plain and simple, without the things that make life comfortable and pleasant
LDOCE: spartan conditions or ways of living are simple and without any comfort

Only one sense can be found in MED and LDOCE for "spartan", so if we use the criterion of MIP, "spartan" in [7] is not metaphorically used. However, as a culture-loaded word, its cultural connotation makes it a direct metaphor, and most researchers take the origin or cultural background information of culture-loaded words as their basic meanings (Dorst & Kaal, 2012; Schmitt 2005). In this case, LDOCE in its CD-ROM provides the etymology information of "spartan" as follows: "of Sparta (16-21 century) from Sparta city in ancient Greece whose people lived simply", which is more than enough for researchers to decide on it metaphorical nature.

2.3 Use of collocation dictionaries

Similar to learner's dictionaries, most contemporary collocation dictionaries are compiled on the basis of large, contemporary, general corpora, and *Macmillan Collocation Dictionary* (henceforth MCD), makes a good choice for identifying metaphors. As one of the most distinguished collocation dictionaries with its unique structure, MCD chooses the high frequent collocations, often associated with the metaphorical meanings of headwords rather their basic meanings, offering help to identify metaphors from following perspectives:

1) selection of headwords

Unlike learners' dictionaries, MCD only includes nouns, verbs and adjectives as headwords, among which, nouns account for 55%, verbs 21% and adjectives and 24% respectively (Coffery 2010). According to Pragglejaz Group (2007), one of the advantages of using dictionaries for metaphor identification is that dictionaries are especially useful for distinguish metaphorical content words from non-metaphorical ones, and for functional words, researchers, to a great extent, have to rely on their intuition. Compared to other collocation dictionaries, eg, *Oxford Collocations Dictionary for Students of English* (henceforth OCD), which has a larger collection of entry words, especially functional words, the headwords included in MCD make it a more convenient means in metaphor identification.

2) segmentation of senses

One of the most distinguished features of MCD is that it highlights metaphorical meanings of lexical words, and in some cases, lists only high frequency metaphorical meanings. Take "cultivate" as an example: both MED and LDOCE have four different senses, with the first two related to concrete senses and last two abstract senses. In OCD, a traditional collocation dictionary, the compliers provide "cultivate + adv" collocation patterns related to three semantic fields (1 land ; 2. crop ; 3. try to develop), while MCD only lists one metaphorical sense for collocations: "develop an attitude, ability, or relationship". Actually, in the entry list of MCD we can find a large quantity of headwords with only metaphorical sense, including "gulf", "ignite" and "veil" etc. and the heightened awareness on metaphor in MCD offers a direct help for researchers to determine the contextual meaning in MIP.

3) choice of collocates

MCD, with collocates based on semantic groups, can help researchers make decisions when use of learners' dictionaries leads to confusion:

[8] *I have to repay $250 every month, and that's a big **chunk** of my salary.*

chunk n.
MED: 1. a large, thick piece of something
2. a large amount of part of something
LDOCE: 1. a large thick piece of something that does not have an even shape
2. a large part or amount of something

Both MED and LDOCE have two meanings, but the infinite pronoun "something", either describes shape of an object in sense 1 or quantity of something in sense 2, is not sufficient to make a judgment about its abstractness, therefore, provides few clues to its metaphorical feature. However, in MCD, the following collocates are listed :

chunk n.
MCD:
a large part or amount of something
- adj+N (ommited)
- N + of food **beef, bread, cheese, chicken, cucumber, lamb, meat, pineapple**
hard solid substance **antonym, ice, masonry, metal, rock, wood**
time or money **budget, day, money, salary, and time**

Although in MCD "chunk" is also defined with "something", its collocates in different semantic groups clear up the confusion caused by the infinity of possibilities in "something", which may blur the distinction between its concrete and abstract senses, and consequently, lead to researchers' frustration in MIP.

2.4 Use of specialized dictionaries

As objectivity and precision are crucial in technical and scientific languages, figurative, vague and ambiguous expressions are, to a great extent, undesirable. What's more, unlike learners' dictionaries that follow descriptive principles, specialized dictionaries are in essence prescriptive. What's more, in contrast to historical dictionaries, they are synchronic rather than chronic, and consequently, specialized dictionaries give little consideration to lexicalization process, in which metaphoricity plays an important part (Temmerman 2000). However, as metaphor is an important vehicle for people to conceptualize the world, not only in daily life, but in all kind of activities, including science, business, and legal activities, the language coded in specialized dictionaries cannot be reduced to literal level. Take business for example, as business language, by its very nature, is metaphorical (Koller 2004; White 2003), figurative expressions will certainly make part of business dictionaries. For instance, data from corpora show that the word "bubble" collocates with words related to business in many cases, however, in MED, and LDOCE neither "bubble" as noun or as a verb relates specifically to business, though we can find its connection to "emotion", "feeling", "activity" and "time" in the given definitions. However, in *Longman Business Dictionary* (henceforth LED), we will find:

bubble n.
LED : 1 when a lot of people buy shares in a company that is financially weak, with the result that the price of the shares becomes much higher than their real value
2 **the bubble bursts** if the bubble bursts in a particular area of business, a period of growth and success ends suddenly

As we shall see, though there is only one sense listed in LED, it's just the metaphorical sense that helps determine contextual meaning more directly, hence more effectively for metaphor identification, especially when we consult LED for a cross reference.

3. Conclusion

Metaphor research is heavily based on metaphor identification, in which MIP is widely applied as a tool. Though in MIP researchers mainly depend on learners' dictionaries to support their intuition, this will be complemented with use of historical dictionaries, collocation dictionaries and specialized dictionaries for a cross reference. There is no denying that dictionary use in metaphor identification is time-consuming, especially with large amounts of data for analysis, and it's less applicable when dealing with functional words, special terms and culture-loaded words, yet compared with introspection and corpus method in metaphor identification, MIP is highly recommended to and universally applied by researchers in metaphor identification for the least dependence on intuition (Zhong & Chen 2013), and it is the use of dictionaries that provides an objective basis for the reliability of the MIP.

References:

Coffey, S.(2011). A new pedagogical dictionary of English Collocations. *Journal of Lexicography*, 24 (3), 328-341.

Diegnan, A. (2005). *Metaphor and corpus linguistics*. Amsterdam/Philadelphia: John Benjamins.

Dorst, A.& Kaal, A. Metaphor in discourse.(2012). In Fiona, M. (Ed). *Metaphor in use: Context, culture, and communication*. Philidelphia: John Benjamins Company, 51-68.

Koller,V. (2004). *Metaphor and Gender in Business Media Discourse: A Critical Cognitive Study*. Palgrave, Basingtoke. UK.

Krennmayr, T. (2013). Top-down versus bottom-up approaches to the identification of metaphor discourse. *Metaphorik.de*, 24, 7-36.

Krennmayr, T. (2008). Using dictionaries in linguistic metaphor identification. In Johansson, N. & D. Minugh.(Eds) *Selected Papers from the 2006 and 2007 Stockholm Metaphor Festivals*. Stockholm: Department of English, Stockholm University, 97-115.

Lakoff, G. & M. Johnson. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse, *Metaphor and Symbol*, 22(1), 1-39.

Steen, G. (2011). The contemporary theory of metaphor- now new and improved. *Review of Cognitive Linguistics*, 9(1), 26-64.

Steen,G., Dorst, A., Herrmann, B., Kaal, A., Krennmayr, T. & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins Company.

Steen, G. (2007). *Finding metaphor in grammar and usage: A methodological analysis of theory and research*. Amsterdam/Philadelphia: John Benjamins Company.

Temmerman, R. (2000). *Towards new ways of terminology secription: The sociocognitive approach*. Amsteredam: John Benjamins.

Zhong Fenglan & Chen Hui. (2013). A review on studies of metaphor identification. *Foreign Language Studies*, 5, 40-44.

White, W. (2003). Metaphor and Economics: The Case of Growth. *English for Specific Purpose*, 22, 131-151.

♭ All examples are from BNC corpus in this paper.

Dictionaries:

Longman Business Dictionary, Pearson Education Ltd, 2007.
Longman Dictionary of Contemporary English, Pearson Education Ltd, 2005.
Macmillan English Dictionary for Advanced Learners, Oxford: Macmillan, 2002.
Macmillan Collocation Dictionary, Oxford: Macmillan, 2010.
Oxford Collocations Dictionary for Students of English, Oxford: OUP, 2002
Oxford English Dictionary: http://www.oed.com/

The Interaction between EFL and English-Japanese Dictionaries[1]

YAMADA Shigeru
Waseda University
shayamda@waseda.jp

Abstract

EFL dictionaries and English-Japanese dictionaries (EJDs) have developed through the interaction with each other. The grading of headwords was initiated by the *Standard EJD* on the basis of Thorndike's and Horn's wordlists and was refined by EFL dictionaries using corpus data. EJDs preceded EFL dictionaries in the indication of stress patters of compounds. Plausibly, the diagram showing sense development in the *Lighthouse EJD 1* inspired the *Macmillan ED*'s provision of menus. *A Grammar of English Words*' indication of verb patterns (through the *Idiomatic and Syntactic ED*) influenced both EFL dictionaries (initially using codes) and EJDs (transparent indications). *Saito's Idiomological EJD* explicitly indicated the combination of selectional restrictions and verb patterns. Corpus-based lexicography has spread from the *COBUILD1* to other EFL dictionaries and EJDs. Although there may have been antecedents and the input from other sources, the interaction between the two genres of learners' dictionaries enhanced their mutual development.

Key words: corpus, EFL dictionary, English-Japanese dictionary, grading of headwords, menu, selectional restriction, signpost, stress pattern, verb pattern

1 Introduction

The history of English-Japanese dictionaries (EJDs) begins in 1862 with the publication of *Ei-wa Taiyaku Shuchin Jisho* (*A Pocket Dictionary of the English and Japanese Language*). Exactly 80 years later in Japan, the first full-fledged EFL dictionary, the *Idiomatic and Syntactic English Dictionary* (*ISED*), was published by Kaitakusha. The dictionary survives today as the *Oxford Advanced Learner's Dictionary* (*OALD*, now 9th ed., 2015). Kihara and Masaoka (1973: 10) state that it is no exaggeration that practically every English-Japanese learner's dictionary which came after the *ISED* was influenced by and benefited from that first dictionary. On the other hand, there is a report that a number of EJDs sat on the bookshelf in the study of A. S. Hornby, one of the editors of the *ISED*. When I visited an EFL dictionary publisher in mid-1990's, I found they also held several EJDs. Tono (2006: 10) points out that many EJDs innovations and features were studied and incorporated into overseas English monolingual learners' dictionaries. Apart from the influence from other sources, it can be said that EFL dictionaries and EJDs have developed through the interaction with each other – direct or indirect, ascertained or unascertained – inspiring, influencing, and benefitting each other, and adapting and adopting each other's useful features for their own purposes.

This paper is an attempt to look at the origin or earliest appearances, the transfer and spread of some important innovations and features that may come from the contact between the two genres of dictionaries. The paper focuses on the following six features: the grading of headwords, the indication of stress patterns, menus and signposts, the indication of verb patterns, selectional restrictions, and the use of corpora.

2 Origin and transfer of important features

2.1 Grading of headwords

It is common for EJDs to indicate the relative importance of headwords with asterisks. For instance, *Kenkyusha's New Collegiate EJD 1* (1967) gives two asterisks to 2,074 words to be mastered during junior high and one asterisk to 6,404 words to be learned during senior high (Preface, 3). However, the sources referred to and the criteria for the indication are usually not exactly made clear.

The indication of important headwords goes back to the *Standard EJD* (1929), using numbers (1 to 10). The indication was based on scientific foundations. Editor, Tsuneta Takehara based the numbering on Thorndike's 100,000 words (1921) and referred to Horn (1926) as a supplementary source (Dohi 1999: 54-55). Takehara divided Thorndike's 100,000 words into 10 groups and indicated the most frequent 10,000 words with "1," the next level of 10,000 words with "2" and so on in the margin (Kojima 1999: 415-416).

British-made EFL dictionaries depended on corpus data for the indication of important headwords. The *COBUILD 2* (1995) introduced the five-level Frequency Bands. The most important 14,700 or so words are given black diamonds in the Extra Column, according to their frequency in the Bank of English:

◆◆◆◆◆ c 700 words
◆◆◆◆◇ c 1,200
◆◆◆◇◇ c 1,500
◆◆◇◇◇ c 3,200
◆◇◇◇◇ c 8,100
(Introduction, xiii)

The *LDOCE 3*, based on the Longman Corpus Network, distinguished between written and spoken English and indicated the most frequent 3,000 headwords in three levels in each of the categories. For instance, **reasonable** is indicated with "S1" and "W2" in the margin (the former above the latter), meaning the word is within the most frequently used 1,000 words of spoken English and within the second most frequently used 1,000 words of written English.

The *OALD 7* introduced the Oxford 3000™. The words of the Oxford 3000 are shown in larger type and with a key symbol. The list of words serves as a defining vocabulary and also as a starting point for vocabulary expansion (p. R99). The 3,000 words were selected on three criteria: frequency from the analysis of corpora, range of use in different text types, and familiarity to users of English; the British National Corpus, the Oxford Corpus Collection, and over 70 experts of teaching and language study were also consulted (*ibid.*).

The words included in *the Academic Wordlist* are so indicated in the *LDOCE 5* (2009) and the *OALD 8* (2010).

The *CALD 2* (2005) indicated important words, meanings, and phrases on the bases of corpus data and recommendation of teachers and academic advisors; there are three levels: E (Essential, 4,900 meanings), I (Improver, 3,300), and A (Advanced, 3,700) (Introduction, vii). The *CALD 4* (2013) labeled important words, meanings, and phrases on the basis of the CEFR (Introduction, ix). The *Ace Crown EJD 2* (2013) referred to the CEFR-J Wordlist for its indication of important words (A1: 1,068 words; A2: 1,247; B1: 2,132; B2: 2,306) (Ace Crown Word Rankings, vi).

2.2 Indication of stress patterns

Takebayashi, *et al*. (1975: 109) point out that the indication of stress patterns of compounds was already provided by the *Union EJD* (1972): e.g., bús stop, associátion football (Higashi, *et al*. 1979: 67).

Takebayashi, *et al*. (1975: 109) count among the *OALD 3*'s (1974) remarkable improvements the indication of stress of separate compounds (e.g., *citrus fruit*, *bank clerk*, *Christmas rose*) and fixed phrases. It was only in the 4th edition (1989) that the indication was in principle extended to all compounds and idioms (Takahashi, *et al*. 1992: 78-80).

Higashi, *et al*. (1979: 67) raise the indication of stress shift as a merit of the *LDOCE 1* (1978).

Akasu, *et al*. (1996: 29-30) discuss the merits and demerits of the *CIDE*'s indications of stress patterns:

> *CIDE* indicates stress patterns of all the phrase-type entries (compounds, idioms and phrasal verbs). This is the great merit of *CIDE* over its rivals, since *LDCE*[2] and *COBUILD*[2] do not indicate the stress patterns of the idioms and phrasal verbs at all. The demerit is that in most cases the stress patterns are given in the Phrase Index only and not in the body of the dictionary (Akasu, *et al*. 1996: 29-30).

2.3 Menus and signposts

As a long-standing, major challenge to the user, Scholfield (1996) points out the task of "wading through this [the sheer mass of condensed target language text in monolingual entries] picking out the numbered definitions and checking each one to find the right one." To assist the user in this task, guide words, signposts, short cuts, and menus have been introduced by the following EFL dictionaries.

Table 1 EFL Dictionaries adopting signposts and menus

Guide words	**Short cuts**	**Signposts & Menus**	**Menus**
CIDE (1995)		*LDOCE 3* (1995)	
	OALD 6 (2000)	*LDOCE 4* (2003)	*MED 1* (2002)
CALD 2 (2005)	*OALD 7* (2005)		*MED 2* (2006)
CALD 3 (2008)	*OALD 8* (2010)	*LDOCE 5* (2008)	

(Taken from Yamada [2013: 199])

Signposting was also incorporated by a native speaker's monolingual dictionary, *Encarta World English Dictionary* (1999), and an EJD, *Progressive EJD 5* (2012)[2].

In connection with the adoption of menus in his *MED 1*, Michael Rundell, editor-in-chief (personal communication) referred to the influence of EJDs and cited the following two reasons for choosing menus over signposts: (1) With the information all at the top of the entry, it is easier to see the full picture; (2) Since the layout of the menus usually allows lexicographers a little more space than is available for signposts, the clues for users are a little more likely to be helpful (Yamada 2010: 164).

An EJD that influenced Rundell's choice is supposed to have been the *Lighthouse EJD 1*. The dictionary indicated the sense development of an important word. The feature was intended not so much to help navigate a polysemous entry as to make clear the semantic and derivational relationship between scattered senses. The one for **head** takes the form of a diagram[3]:

2.4 The indication of verb patterns

The *GEW* is credited with introducing a systematic indication of verb patterns by means of codes which were "later to be applied, with minor or major variations, in the first four editions of *ALD* and in various rival compilations" (Cowie 1999: 37) (see Table 2). In the *LDOCE 1*, alpha-numeric codes were extended to nouns and adjectives. In early 1980s, however, the gap between some sophisticated design features of EFL dictionaries and the users' rudimentary reference skills was pointed out (Cowie 1981: 206), including grammar codes (Bejóint 1981: 16, 19). In reaction to academic reviews and the publisher's international user research the *LDOCE 2* turned to transparent indications, abandoning the codes and abbreviations. The dictionary placed a transparent verb pattern before an example (see Table 2).

Table 2 Indications of "*want*+object+*to*-infinitive"

GEW	*OALD*	*LDOCE*
V. P. 17.	*ISED* (1942) *vt. & i.* ❷ (P3)	
	ALD2 (1963) *v.t. & i.*	

	2. (VP3)	
	3/e (1974) *vt*, *vi* **2** [VP17]	1/e (1978) *v* [Wv6] 1 [… V3 …]
	4/e (1989) *v* **1** [… Tnt no passive …]	2/e (1987) *v* [*not usu. in progressive forms*] **1** [T] … [+*obj*+*to-v*] *He wants you to wait here.*

(Adapted from Yamada [2013: 193])

The *ISED*'s indication of verb patterns influenced EJDs as well. Unlike the first generations of EFL dictionaries, the bilingual dictionaries indicated the patterns in accessible ways. For example, *Kenkyusha's New Collegiate EJD 1* (1967)[4] provided the sentence pattern 〔+object+to do/+object+doing〕 before translational equivalents:

> **want** … *vt*. **1** … **d**〔+目+*to* do/+目+do*ing*〕〈…に…することを〉望む, 〈…に…して〉ほしいと思う：She ～*s* me *to* go with her. 彼女は私に一緒に行ってもらいたがっている … (emphasis added)

The *Lighthouse EJD 1* (1984) gave a similar sentence pattern to the example, the pattern being inserted between the example and its Japanese translation.

2.5 Selectional restrictions

It has been a customary practice for EFL dictionaries to mark off selection restrictions in parentheses in the definition. In a full-sentence definition, they are incorporated in the subordinate clause.

Published as far back as in 1915, *Jukugo Honi Ei-wa Jiten* (*Saito's Idiomological EJD*) is remembered as a landmark and is still praised for its innovative features. It is noteworthy that the dictionary indicated the combination of selectional restrictions and verb patterns in front of translational equivalents (Takebayashi 1992: 506, Kojima 1999: 386-7):

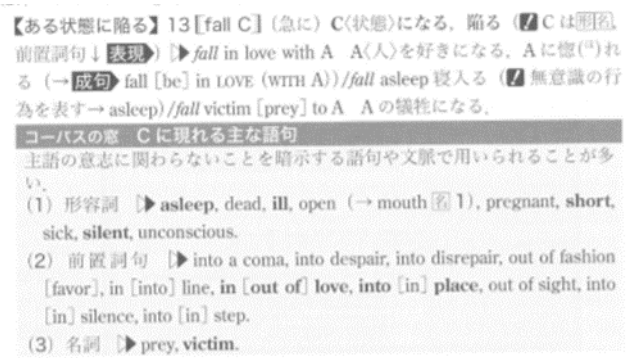
> **Ob-ject'** … **【自動】**（Consentに對し— **to** some plan）…
> [【intransitive】(opp. *consent*, ～ **to** some plan) …]
> **【他動】**（something **to** or **against** a statement, theory, etc）…
> [【transitive】…]

2.6 The use of corpora

After the publication of *COBUILD 1* (1987), corpus basis became standard in the compilation of EFL dictionaries. EJDs greatly lagged behind EFL dictionaries in the use of corpus data (Tono 2006: 11). It was only in 2003 that the first corpus-based dictionary, *Wisdom EJD 1*, was published in Japan. The project team developed a 100 mil.-word balanced corpus, Sanseido Corpus, for the use for their dictionary.

EJDs have recently made remarkable progress in the application of corpus data for Japanese students of English. Co-Editor-in-Chief of the *Wisdom EJD*, Nagayuki Inoue (2016: 31) recalls that they analyzed corpora so closely as to obtain the kind of information Japanese students would find useful. Corpus-informed usage notes have been introduced since its second edition (2006) to help users' production in English (*ibid*. 33). For example, the one at sense 13 of **fall** in the 3rd edition (2012) points out that the verb is used with words and

phrases or in contexts that imply the non-initiative of the subject, listing abundant complementation patterns of (1) adjectives, (2) prepositional phrases, and (3) nouns.


【ある状態に陥る】13 [fall C] (急に) C〈状態〉になる，陥る (/ C は形名，前置詞句↓ 表現) [▶ *fall* in love with A　A〈人〉を好きになる，A に惚(ほ)れる (→ 成句 fall [be] in LOVE (WITH A)) /*fall* asleep 寝入る (/ 無意識の行為を表す→ asleep) /*fall* victim [prey] to A　A の犠牲になる．

コーパスの意　C に現れる主な語句

主語の意志に関わらないことを暗示する語句や文脈で用いられることが多い．

(1) 形容詞 [▶ **asleep**, dead, **ill**, open (→ mouth 名 1), pregnant, **short**, sick, **silent**, unconscious.

(2) 前置詞句 [▶ into a coma, into despair, into disrepair, out of fashion [favor], in [into] line, **in** [**out of**] **love**, **into** [in] **place**, out of sight, into [in] silence, into [in] step.

(3) 名詞 [▶ prey, **victim**.

(Taken from Inoue [2016: 37])

Information of this detail cannot be found in any other work: EFL dictionaries (the *Macmillan English Dictionary 2*) or grammars (Huddleston and Pullum [2002], Swan [2005], Quirk, *et al*. [1985]) (Inoue 2016: 33-35).

The *Progressive EJD 5* (2012), intended for adults and business people, is based on a 2.1 bil.-word SEKAI Corpus. It includes the British National Corpus (BNC) and PERC [Professional English Research Consortium] Corpus, ranging from economics, law, politics to computer (Tono 2016: 52-54). There are innovative corpus-based features. One is the notes which contrast general and field-specific collocations. For example, the note at **figure** treats 一般 ('general') vs. 経済 ('economics') collocations. It also reminds the user that *figure* tends to be used in the sense of 'person' in the former and 'number' in the latter:


コーパス 形＋figure

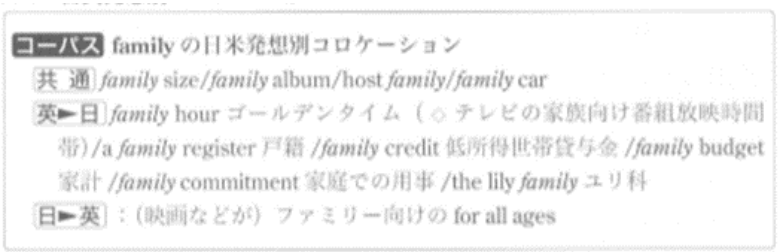
一般 public/key/leading/political/late/senior/official/double

経済 big/net/low/high/full/average/representative/animated

◆ 一般 では「人物」(⇨3) の意味が多いが，経済 では「数字」(⇨2) が多い．

(Taken from Tono [2016: 55])

Another is the notes dealing with collocations, reflecting the difference in the ways of thinking between English and Japanese. For instance, the one at **family** lists collocates under three categories: 共　通 ('common'), 英▸日 ('English to Japanese'), and 日▸英 ('Japanese to English'). The last category indicates the Japanese collocation that includes the loan word ファミリー (the transliteration of *family*) with its English translation:


コーパス family の日米発想別コロケーション

共 通 *family* size/*family* album/host *family*/*family* car

英►日 *family* hour ゴールデンタイム (◇ テレビの家族向け番組放映時間帯) /a *family* register 戸籍 /*family* credit 低所得世帯貸与金 /*family* budget 家計 /*family* commitment 家庭での用事 /the lily *family* ユリ科

日►英 : (映画などが) ファミリー向けの for all ages

(*ibid*. 56)

3 Conclusion

It is difficult to trace the origin and spread of features of learners' dictionaries even among EFL dictionaries and EJDs. However, this paper yields the following findings. The grading of headwords was initiated by the *Standard EJD* (1929) based on Thorndike's and Horn's wordlists; EFL dictionaries after 1995 provided headwords with corpus-based frequency information. In the indication of stress patters of compounds, EJDs preceded EFL dictionaries. It can be assumed that the indication of sense development in the *Lighthouse EJD 1* (1984) inspired the *MED*'s provision of menus. The *ISED*'s indication of verb patterns influenced those in EJDs. It is plausible that the user-friendly indication of verb patterns in EJDs in turn spread to EFL dictionaries. An early indication of selectional restrictions can be found in *Saito's Idiomological EJD* (1915), which explicitly indicated the combination of selectional restrictions and verb patterns. The use of corpora in dictionary compilation has spread from the *COBUILD1* (1987) to other EFL dictionaries and EJDs. Although the origin and developments may have had antecedents or been inspired by other sources outside the scope of this paper, it can safely be concluded that there was some interaction between the two genres of learners' dictionaries, which were good for their mutual development. Although conceivable innovations may have been exhausted in EJDs (Inoue 2016: 31) and also in EFL dictionaries, interaction will continue to stimulate and benefit each type in the description and presentation of the common core and specific areas of the English language.

Notes

1 I would like to express my gratitude to Professor Jeffrey Miller for his help with the final draft.

2 The "signposts" in the *Progressive EJD 5* serve also as the indicators of sense groups (Preface). This idea and devise go back to the *Global ED* (1983) (cf. Nakao 1989: 298).

3 The numerals correspond to the sense numbers.

4 In the preface to *Kenkyusha's New Collegiate EJD2* (1968), the following works are credited with thanks: *ISED* (under the new title of *The Advanced Learner's Dictionary of Current English* [new ed., 1963] and Hornby's *A Guide to Patterns and Usage in English*.

References

Akasu Kaoru, *et al*. 1996. "An Analysis of *Cambridge International Dictionary of English*." *Lexicon*. No. 26. Tokyo: Iwasaki Linguistic Circle. 3-76

Béjoint, Henri. 1981. "The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills." *Applied Linguistics* 2. 207-222.

Cowie, A. P. 1981. "Lexicography and its Pedagogic Applications: An Introduction." *Applied Linguistics* 2. 203-206

Cowie, A. P. 1999. *English Dictionaries for Foreign Learners: A History*. Oxford: OUP.

Dohi, Kazuo. 1999. "Thondike to Amerika no Gakushu Jiten" [Thorndike and Learner's Dictionaries in the U. S. A.]. *Toyoko English Studies*. No. 8. Tokyo: Toyoko Gakuen Women's College. 17-77.

Higashi, Nobuyuki, *et al*. 1979. "*Longman Dictionary of Contemporary English* no Bunseki (Paato 1)" [An Analysis of *Longman Dictionary of Contemporary English* (Part 1)]. *Lexicon*. No. 8. Tokyo: Iwasaki Linguistic Circle. 45-101.

Horn, Earnest. 1926. *A Basic Writing Vocabulary: 10,000 Words Most Commonly Used in Writing*. University of Iowa.

Inoue, Nagayuki. 2016. "Wizudamu Ei-wa Jiten" [Wisdom English-Japanese Dictionary]. Eds. Minamide, Kosei, *et al*. 21-40.

Kihara, Kenzo, and Keiko Masaoka. 1973. "Gakushu Ei-wa Jiten no Nagare" [The Recent Trends in English-Japanese Learners' Dictionaries]. *Gendai Eigo Kyoiku* (*Modern English Teaching*). Vol. 9. Tokyo: Taishukan. 10-11.

Kojima, Yoshiro. 1999. *Eigo Jisho no Hensen: Ei, Bei, Nichi wo Awasemite* [The Development of English Dictionaries: A Parallel Look at Britain, America, and Japan]. Tokyo: Kenkyusha.

Minamide, Kosei, *et al*., eds. 2016. Eigo Jisho wo Tsukuru (English Dictionary Making: Practice and Research). Tokyo: Taishukan.

Nakao, Keisuke. 1989. "English-Japanese Learners' Dictionaries." *International Journal of Lexicography* Vol. 2. No. 4. 295-314.

Nakao, Keisuke. 1998. "The State of Bilingual Lexicography in Japan: Learners' English-Japanese/Japanese-English Dictionaries." *International Journal of Lexicography* Vol. 11. No. 1. 35-50.

Scholfield, P. 1996. "Why Shouldn't Monolingual Dictionaries be as Easy to Use as Bilingual Ones?" *Longman Language Review*. Issue Number Two.

Takahashi, Kiyoshi, *et al*. 1992. "An Analysis of *Oxford Advanced Learner's Dictionary of Current English*, Fourth Edition." *Lexicon*. No. 22. Tokyo: Iwasaki Linguistic Circle. 59-200.

Takebayashi, Shigeru, *et al*. 1975. "*Oxford Advanced Learner's Dictionary* no Bunseki" [An Analysis of *Oxford Advanced Learner's Dictionary*]. *Lexicon*. No. 4. Tokyo: Iwasaki Linguistic Circle. 68-114.

Takebayashi, Shigeru. 1992. "Ei-wa Jiten" [EJDs]. Eds. Takebayashi, Shigeru, *et al*. Sekai no Jisho [Dictionaries in the World]. Tokyo: Kenkyusha. 505-531.

Thorndike, Edward. 1921. *The Teacher's Word Book*. Teachers College, Columbia University.

Tono, Yukio. 2006. "Learner's Dictionary Gaikan" [An Overview of Learners' Dictionaries]. *Nihongogaku*. Vol. 25. Tokyo: Meiji Shoin. 6-20.

Tono, Yukio. 2016. "Deta Saiensu to Yuza no Tetsugaku" [Data Science and User Philosophy]. Eds. Minamide, Kosei, *et al*. 41-58.

Yamada, Shigeru. 2010. "EFL dictionary evolution: Innovations and drawbacks." *English Learners' Dictionaries at the DSNA*. Eds. Kernerman, Ilan J., and Paul Bogaards. Tel Aviv: K Dictionaries. 147-168.

Yamada, Shigeru. 2013. "Monolingual Learners' Dictionaries – Where Now?" Ed. Jackson, Howard. *The Bloomsbury Companion to Lexicography*. Ch. 4.5. London: Bloomsbury. 188-212.

On the Inclusion of New Words in *A New English-Chinese Dictionary*

Yongwei Gao
Fudan University
ywgao@fudan.edu.cn

Abstract

A New English-Chinese Dictionary (NECD), one of China's best-selling bilingual dictionaries, was first published in 1976 and so far has undergone three revisions. Its fourth edition, published in 2009, was well received by dictionary critics (Chen 2009; Wang 2010; Xu 2011) and general readers alike. More specifically, the dictionary was much lauded for its effort in recording the latest additions to the English vocabulary. Chen (2009:241) believes that "the NECD4 is a successful bilingual dictionary which presents a panorama of contemporary English lexicon". With a massive makeover (e.g. separate entries for niched compound words and reordering of senses) and an addition of more than 5,000 neologisms, the fourth edition has, to some extent, withstood fierce competition from the bilingualized editions of the Big Four which have dominated the English-Chinese dictionary market since the late 1990s. However, as the archetype of a bilingual dictionary compiled by Chinese scholars, the NECD should always stay ahead of the curve in terms of new-word inclusion despite the fact that it is in essence a medium-sized dictionary. This will mean that the compilation team behind the on-going revision of the dictionary should spare no effort in recording new English words that cropped up or became popular in the past decade. This paper attempts to examine the strategies the revisers of the NECD4 should adopt in adding new entries to the dictionary, which include more focus on technical terms, the exclusion of buzzwords, more inclusion of World English words, more coverage of new words that have spawned derivatives, etc.

Keywords: *A New English-Chinese Dictionary*, revision, new words

Introduction

It was conservatively estimated that the number of new words that appear in the English language each year reached 800 (Landau 202). But as the estimation was made in the pre-digital days, this number was by all means an underestimation. Thanks to technological advances in recent decades and the easier access to the Internet in particular, new words are coined more often than ever before, and they are circulated much faster and more widely. The fact that the editorial team of the *Oxford English Dictionary* includes approximately 2,000 new words each year through quarterly updates is a clear indication that the English language now boasts more new words than ever before. No matter what the actual number is, there is broad consensus among dictionary-makers worldwide that they now have more new-word entries to edit than in the past. With people's increasing awareness of and interest in new lexical kids on the block, dictionary editors, monolingual or bilingual, are sparing no efforts in recording as many neologisms as possible in their dictionaries, and the advertising of some select new terms in the blurbs or through other means (e.g. tweets, blogs, newspaper articles) has become the norm.

With so many new words on hand, dictionary editors are spoilt for choice. Then what kinds of neologisms should lexicographers record in their dictionaries? Much ink has been spilled over the discussion of the selection criteria for new-word entries in dictionaries (Agnes 1995; Sheidlower 1995; Barnhart 2007). Allan Metcalf (2002) suggests the FUDGE factors in identifying new words, namely frequency of use, unobtrusiveness, diversity of users and situation, generation of other forms and meanings, endurance of the concept. In reality, although an unpredictable mixture of these factors has been used to decide which terms that might be able to stand the test of time, different dictionary editors may rely on their own set of criteria, one of which is frequency of use which is usually indicated with the help of their respective corpus or corpora.

Since its publication nine years ago, the fourth edition of *A New English-Chinese Dictionary* (NECD) has been favorably received by dictionary critics (Chen 2009; Wang 2010; Xu 2011) and general readers alike. More specifically, the dictionary was much lauded for its effort in recording the latest additions to the English vocabulary. Chen (2009:241) believes that "the NECD4 is a successful bilingual dictionary which presents a panorama of contemporary English lexicon". With a massive makeover (e.g. separate entries for niched compound words and reordering of senses) and an addition of more than 5,000 neologisms, the fourth edition has, to some extent, withstood fierce competition from the bilingualized editions of the Big Four which have dominated the English-Chinese dictionary market since the late 1990s. However, as the archetype of a bilingual dictionary compiled by Chinese scholars, the NECD should always stay ahead of the curve in terms of new-word inclusion despite the fact that it is in essence a medium-sized dictionary. This will mean that the compilation team behind the on-going revision of the dictionary should go out of its way to record new English words that cropped up or became popular in the past decade. This paper attempts to examine the strategies the revisers of the NECD4 should adopt in adding new entries to the dictionary, and the author, as editor-in-chief of the NECD5, will set out several criteria include more coverage of new words that have spawned derivatives, more focus on technical terms, more inclusion of World English words, the exclusion of buzzwords, etc.

1. A Brief History

In the first two decades since the founding of the People of Republic of China in 1949, no medium- or large-sized English-Chinese dictionaries were ever compiled in the country. It wasn't until the mid-1970s that saw the publication of a medium-sized bilingual dictionary

that has almost become synonymous with English-Chinese lexicography, namely the NECD. So far the NECD has undergone three revisions.

1.1 The first edition

Boasting 80,000 entries, the first edition was published in 1975 after more than five years of compilation. As a lexicographical product created during the Cultural Revolution, the first edition of the NECD was riddled with ideological preaching (as in illustrative examples such as "In 1949 New China, like a sun rising in the east, *appeared* in the world" and "The Red Guard *aspires* to be a PLA fighter"), but it achieved unexpected successes at home and abroad. Shortly after its publication, it received rave reviews in several American newspapers. *The New York Times*, for instance, carried an article entitled "New Chinese Dictionary Tells It Like It Is" on Jan. 22, 1976, commending its effort in recording "current American idioms and scientific and technical vocabulary".

The NECD compilation team, headed by the late professors Ge Chuangui and Lu Gusun, followed the footsteps of previous English-Chinese dictionary-makers in recording as many commonly used words as possible. Some of the new words they recorded in the dictionary include *acid* (in the sense of LSD), *antinovel*, *cruise missile*, *graviton*, *machine language*, *skyjack*, *streak* (to run naked through a public space), *with-it*, etc.

1.2 The second edition

The second edition of the NECD was published in 1985 after about six hundred changes had been made to the definitions and illustrative examples and an addendum of about 4,000 new-word entries was provided after the A-to-Z part in the dictionary. The addendum was put together on the basis of several existing English dictionaries such as *Barnhart Dictionary Companion*, *Concise Oxford Dictionary* (7th Edition), and *The Supplement to the OED*. The 4,000 entries it recorded were mostly new words, and a small proportion of entries were existing words with new meanings or usages. The then new acronym *AIDS* was one of them. Its inclusion exemplified the NECD revisers' effort its compilers made in recording the latest English vocabulary given the fact that the disease was first identified three years prior to the publication of the dictionary. Other new words that made the cut include *aerobics, big bang theory, break dancing, hedge fund, hopefully, Turner's syndrome, underwhelm, videoconference,* etc.

1.3 The third edition

The third edition is more popularly known as the century edition since it was published in the year 2000. The revision took about six years intermittently. In this edition, thousands of new words were included along with hundreds of new senses. Such new additions, combined with other inclusions, pushed the number of total entries to about 100,000. Wu Ying, its editor-in-chief, wrote that "the century edition made the inclusion of those English new words, new meanings, and new usages that appeared in the past two or three decades a top priority of the revision"(Wu 2000: Preface). Ding Zhicong (2001: 118) compared the NECD's inclusion of technical terms with that of *Far East English-Chinese Dictionary* and the eighth and ninth editions of *Merriam-Webster's Collegiate Dictionary*, pointing out that the NECD3 included far more terms than its counterparts. As a matter of fact, the revisers did add hundreds of technical terms to this edition. Wu singled out dozens of technical entries for discussion (Wu 2001: 120), including neologisms such as *desktop, Ebola virus, genome, mobile phone, palmtop, and server,* along with several *-ware* combinations (e.g. *groupware, kidware, freeware, wetware*). As one of the revisers was previously involved in the compilation of *The Supplement to The English-Chinese Dictionary* (published in 1999) which

was virtually a new-word dictionary containing about 3,500 entries, many of the NECD3's new entries were based almost verbatim on the supplement, such as *angioplasty, central banker, e-commerce, glass ceiling, global warming, karoshi, mad cow disease*, and *virtual reality.*

1.4 The fourth edition

It took the editorial team more than two years to revise the third edition. The revision, rather extensive in scale, attempted to make the dictionary more user-friendly through the addition of learner-oriented features (e. g. more illustrative examples and the addition of usage notes) along with routine revisions such as the correction of errors (mostly misspellings and typos), the inclusion of new words and new senses, and the improvement of Chinese translation of either definitions or illustrative examples. The dictionary made headlines nationwide when it came out in 2009. What was widely reported about the dictionary, however, is not its shift towards a learner-oriented dictionary, but its wide coverage of the latest English lexicon. Most widely used English new words and expressions popularized in the first few years of the 21st century were recorded in the fourth edition, such as *bromance, carbon neutrality, chatbot*[1]*, helicopter parent, microfinance, recessionista, smishing, tweet,* and *canary in a coal mine*. What distinguished the revisers of the fourth edition from their predecessors is that the former made their selection of new words from their own research rather than based their selection on English dictionaries then available.

1.5 The fifth edition

The ongoing revision of the NECD was started late last year, and this new edition is expected to come out in 2019. The revision will involve the clearing of lexicographical cobwebs through deleting obsolete or archaic terms and meanings (e.g. *abiosis, Achromycin, adhibit, admeasurement, aedile, aerography, aestival, affranchise*), the amendment of Chinese equivalents (e.g. those of technical terms in particular), and the furnishing of more illustrative examples for common words, etc. And most important of all, the revisers have set a goal of entering at least three thousand new words that will encompass social, political, cultural, technological aspects of human life. In order to reach a more balanced inclusion of headwords, it is of great necessity to discuss the criteria to be adopted.

2 Inclusion of new words in the NECD5

Dictionary revision usually involves the addition of new words and/or illustrative examples along with amendments to the microstructure of an entry. In the bilingual context, lexicographical revision may also involve the amendment of existing equivalents. When it comes to the inclusion of new words in English-Chinese dictionaries, the criteria involved have been discussed by a few scholars including Thomas Creamer. He Creamer 102) not only advocated two main principles for the selection of headwords, one of which is "is that the lexical items should be general-language terms that appear in current English-language periodicals" (, but also discussed six selection criteria that include "The word is a common phrase or expression that may confuse a nonnative speaker of English", "It is a popular scientific term", and "It is an established word that has acquired a new meaning". But as Creamer based his discussions on the basis of the compilation of the above-mentioned

[1] Surprisingly enough, this word was one of the Word of the Year shortlist choices of Oxford Dictionaries in 2016.

supplement to the ECD, not all his criteria are relevant in the revision of the NECD. Instead, the revisers, while carrying on the tradition of their predecessors, will adopt a multi-pronged approach to the inclusion of new-word entries which centers around one main principle and several sub-principles. The main principle, rather general in nature, refers to the inclusion of neologisms that have come into common parlance in the past one or two decades. In this regard, two criteria will be used to judge whether they are common or not:

A. They should frequently appear in newspapers and magazines or online news sites in the English-speaking countries. *Bitcoin*, referring to a decentralized digital currency, is a case in point. As a search of this word through the *News on the Web* (NOW) corpus (with more than 6 billion words) comes up with 93,144 and 11,681 hits respectively for its singular and plural form (as of May 12, 2018), such a frequency merits its inclusion into the dictionary, and so do its related terms such as *blockchain*, *cryptocurrency*, and *initial coin offering* (ICO). Similarly, *crowdfunding*, a lexical creation based on *crowdsourcing*, appeared 23,397 times in the above-mentioned corpus, so it will be entered in the NECD5 along with *crowdfunder* and its back-formation *crowdfund*.

B. They should be able to generate derivatives or related words. *Alternative right*, for example, popularized two years ago during the U.S. general election, has become a strong candidate for inclusion as it has already spawned a string of related words such as *alternative left, alt-right, alt-left, alt-light, alt-righter*, and *alt-rightist*. Another typical example is *Brexit*, a combination of *Britain* and *exit*. Although it is in essence a topical word, given the fact that it is exerting and will exert great influences upon the British society, its inclusion into any dictionary is understandable. As a matter of fact, the OED included the word during its quarterly update in March, 2017. So far *Brexit* has given rise to several derivatives and related words such as *Brexiteer*, *Brexiter*, *Bremain*, and *regrexit*. Therefore, its inclusion into the NECD5 is a foregone conclusion.

As regards the sub-principles, they refer to the inclusion of more technical terms, the inclusion of more blend words, the inclusion of more World English words, the avoidance of predictable formations and buzzwords.

2.1 The inclusion of more technical terms

As is often the case with new additions in English dictionaries, a large proportion of them are technical terms, which, of course, can be attributed to technological advances. Because many of these terms are no longer restricted to professionals, their inclusion in the NECD5 makes great sense and will ultimately benefit its users.

Let's take medical terms for example. The NECD5 plans to record at least one hundred new medical terms. One salient feature of these terms will be the inclusion of more initialisms such as *APS* (antiphospholipid syndrome), *ASD* (autistic spectrum disorder), *CAPS* (catastrophic antiphospholipid syndrome), *FNA* (fine needle aspiration), *HDV* (hepatitis D virus), *IUS* (intrauterine system), *LVAD* (left ventricular assist device), and *NSSI* (non-suicidal self injury). Meanwhile, several acronyms will also be recorded, such as *MERS* (Middle East respiratory syndrome), *NOTES* (natural orifice translumenal endoscopic surgery), *TURP* (transurethral resection of the prostate), and *SADS* (sudden adult death syndrome). Meanwhile, the compilers of the NECD5 will also include the following three types of medical terms:

A. Words denoting diseases or viruses such as *antiphospholipid syndrome, autism spectrum disorder, chikungunya, diabesity, compartment syndrome, Couvade syndrome, Hughes*

syndrome, *insulin resistance syndrome*, *lifestyle disease*, *norovirus*, and *Zika virus*.

B. Terms describing medical conditions, symptoms or techniques such as *anaphylactic shock, fecal occult blood, fine needle aspiration, hymenoplasty, photoablation, virotherapy*, and *virtopsy*.

C. Miscellaneous terms such as *assisted dying, bikini medicine, concierge medicine, integrative medicine***,** *palliative care, patient zero, postcode prescribing, savior sibling, slow medicine, telehealth*, and *the worried well***.**

Besides medical terms, the NECD5 will also enter terms used in a wide range of subject fields, such as biochemistry (e.g. *adipokine* and *kisspeptin*), biology (e.g. *CRISPR* and *metagenome*), botany (e.g. *aronia* and *lucuma*), chemistry (e.g. *benzylpiperazine* and *copernicium*), computing (e.g. *ad blocker* and *GPU*), economics (*deleveraging* and *double dip*), psychology (e.g. *emotional intelligence* and *size-weight illusion*), and telecommunications (e.g. *femtocell* and *IMEI*).

2.2 The inclusion of more blend words

According to John Ayto, blending "was establishing itself at the end of the 19th century, and the 20th century has taken to it with great enthusiasm" (Ayto xii). Since the turn of the century, there has been a growing tendency to create new words through blending. Although most of such portmanteau words are used in informal contexts, many of them have made their way into dictionaries thanks to their usefulness in expressing combined concepts or ideas. Blending used to be regarded as a minor word-formation process. In Cannon's study of 13,683 neologisms, blending only accounted for a meager 1%. Algeo found that that blending accounted for 5% of all the new words appeared in the column "Among the New Words" in *American Speech* (Algeo 14). In the NECD4, the percentage of blend words greatly increased, almost reaching 8% of all the new words recorded. These new portmanteau words include *bridezilla* (bride + Godzilla), *bullycide* (bully+suicide), *cenbank* (central + bank), *dykon* (dyke + icon), *floordrobe (*floor + wardrobe), *globesity* (global + obesity), *lamestream* (lame + mainstream), *mobisode* (mobile + episode), etc. As more new blends were created in the past decade, the NECD5 will continue to record as many of them as possible. However, as some of the new blends were created for jocular or humorous purposes and usually do not enjoy longevity, they will definitely be excluded, such as *bagnut* (bagel + doughnut, as is popularized by *Orange Is the New Black*), *dadiot* (dad + idiot), *dormcest* (dorm + incest), *mancipation* (man + emancipation), *shress* (shirt + dress), and *textruption* (text + interruption).

What the NECD5 intends to record is dozens of popularly used blends that can be roughly classified into the following three types, as is illustrated in Table One:

Blends with second element clipped	Blends with first element clipped	Blends with both elements clipped
banjolele (banjo + ukulele), *bashtag* (bash + hashtag), *clicktivism* (click + activism), *decacorn* (deca- + unicorn), *droneport* (drone +airport), *hackerazzi* (hacker + paparazzi), *mumpreneur* (mum + entrepreneur), *mantyhose* (man + pantyhose)	*athevening* (athletic + evening), *bleisure* (business + leisure), ***cannabusiness*** **(**cannabis + business), *chemsex* (chemical + sex), *gastroporn* (gastronomy + porn), *hangry* (hungry + angry), *symlink*	*Chinglish* (Chinese + English), *churnalism* (churn out + journalism), *dorgi* (dachshund + corgi), *glocal* (global + local), *jeggings* (jeans + leggings), *subvertising* (subvert +advertising), *sysadmin* (system +

	(symbolic + link)	administrator)

Table One Three types of blend words

2.3 The inclusion of more World English words

As the concept of World English gradually gains traction in the dictionary-making scene, more words from other varieties of English have been included in general monolingual and bilingual dictionaries. Take Australian English for example. *The English-Chinese Dictionary* recorded several hundred Australianisms in its second edition in 2007, including colloquialisms such as *barbie* (barbeque), *brekkie* (breakfast), *coldie* (a cold bottle of beer), *mozzie* (mosquito), *servo* (petrol station), and *waxhead* (someone who surfs on waves). However, the NECD4 did a poor job in including World English words as only four new Australian words were recorded, namely *Australian salute* (the waving of one's hand in front of the face at regular intervals in order to prevent flies from landing on it), *barbecue-stopper* (a topic of conversation that is very interesting or controversial), *grey nomad* (any elderly retired person who spends time travelling around the country in a mobile home), and *silver beet* (a variety of beet).

In view of the fact that Chinese students studying in Australia account for a large proportion of foreign students there and Mandarin has now become the second-most spoken language in Australia[2], the addition of more Australianisms will definitely be a great help in the course of cultural exchanges. As a result, the NECD5 is expected to include dozens of Australians, as is exemplified in Table Two:

Entries	Definitions
chew-'n'-spew	any fast-food restaurant considered to be serving poor quality food
crash-hot	extremely impressive
micro-party	a small political party, esp. one focusing on a single issue
pash rash	an inflammation of the skin caused by passionate kissing with a man with a stubbly face
rev-head	a motor-sport enthusiast
share plate	a serving or selection of food to be shared between two or more diners
shirt-front	to confront in a threatening manner

Table Two Australian English words to be included

[2] The fact that *daigou* (代购), a Chinese borrowing in Australian English meaning "a person outside China who purchases goods for customers in mainland China", was shortlisted for Macquarie Dictionary Word of the Year in 2017 speaks volumes.

However, some World English words that may not be very relevant in the Chinese context, such as *bunny chow* (curry served in a hollowed-out loaf of bread, from South African English), *carnap* (to steal a car, from Philippine English), *condotel* (condominium + hotel, from Southeast Asian English), *keema* (minced meat, from Indian English), *multi-starrer* (a film having an ensemble cast featuring many star performers, from Indian English), *tenderpreneur* (tender and entrepreneur, from South African English), and *timepass* (from Indian English), etc. These words will not be considered candidates for inclusion.

2.4 The avoidance of colloquial or slang words formed by popular (would-be) combining forms

The past few decades have witnessed the creation of many combining forms in the English language that are very productive in forming new words, such as *-core* (as in *nerdcore*), *-ista* (as in *fashionista*), *-licious* (as in *bootylicious*), *-mageddon* or *-geddon* (as in *Snowmageddon*), *-pocalypse* (as in *airpocalypse*), *-tacular* (as in *craptacular*), and *-tastic* (as in *abtastic*). However, as many of their combinations are to a great extent colloquial or slang expressions, they may not deserve a place in an English monolingual dictionary, let alone a bilingual one. Let's take *anorexia* for example. Although *-(o)rexia* has not been officially recognized as a combining form, its suffix-like combinational power has been seen in use, such as *bleachorexia, carborexia, drunkorexia, tanorexia*, and *wannarexia*. The same can be said of *influencer* and its related words. The OED dates this word back as early as 1664, nevertheless, it has not been seen in common use until recently. So far *influencer* has not only formed a widely-used compound word *influencer marketing*[3], but also established itself as a potential combining form, as is attested by the emergence of new combinations such as *fitfluencer, techfluencer, manfluencer*, and *thinkfluencer*. Meanwhile, caution should also be exercised when considering whether the analogical formations in Table Three are suitable for inclusion or not:

Original words	New words patterned after them
Frankenstein	Frankenbite, Frankenburger, Frankenfruit, Frankenstorm, Frankenword[4]
heterosexual	cissexual, lumbersexual, omnisexual, pomosexual, retrosexual, sapiosexual, technosexual, ubersexual[5]
hijacking	brandjacking, crisisjacking, newsjacking, pagejacking, trendjacking[6]
selfie	belfie, dronie,felfie, killfie, shelfie, shoefie
vegetarian	locatarian, meatarian, nutarian, pollotarian, vagitarian[7]

Table Three Analogical formations

[3] This refers to a form of marketing in which focus is placed on influencers—individuals that have influence over potential buyers—rather than the target market as a whole.
[4] *Frankenfish* and *Frankenfood* have already found their way into the NECD4.
[5] *Metrosexual*, the prototypical word in this regard, was recorded in the NECD4.
[6] The still popularly used *clickjacking* has already been included in the NECD4.
[7] Another two *-tarian* words, namely *flexitarian* and *pescetarian*, were included in the NECD4.

In a similar vein, buzzwords should be avoided in the NECD5 as well. A case in point is *fake news* (namely "news stories with false information"). Also popularized during the US general election two years ago, the term was even crowned as The Word of the Year by the American Dialect Society in 2017. Although it is still enjoying wider currency (e.g. 42749 times in the NOW corpus), it is in essence a vague buzzword and should not be recorded in the NECD5. Other faddish words or expressions that should be eschewed include *Antifa* (a political protest movement), *dark data* (data that is unanalyzed, inaccessible, or not organized in such a way that they are readily available), *kompromat* (info collected for use in blackmailing, discrediting, or manipulating someone), *milkshake duck* (someone or something on social media that initially appears endearing but is later discovered to be deeply flawed), *white fragility* (discomfort and defensiveness on the part of a white person when confronted by information about racial inequality and injustice), etc.

Concluding remarks

The revision of dictionaries cannot be divorced from the semantic changes made to existing words, i.e. the addition of new meanings. However, as new meanings are much harder to detect than new words and their lexicographical description is always far from satisfactory, it is impossible to have a systematic discussion of them in this paper. As regards the inclusion criteria of new words, what has been discussed is far from complete as other factors will be also in play in the course of the selection of new words. These are some of the issues that need to be addressed in future studies concerning the selection of neologisms in the revision of bilingual dictionaries.

References

Agnes, Michael. (1995). Why It Isn't There: Practical Constraints on the Recording of Neologisms. *Dictionaries: Journal of the Dictionary Society of North America*, 16, 45-50.

Algeo, John. (1991). *Fifty Years "Among the New Words": A Dictionary of Neologisms, 1941-1991*. Cambridge: Cambridge University Press.

Ayto, John. (2006). *Movers and Shakers: A Chronology of Words That Shaped Our Age*. Oxford University Press.

Barnhart, David K. (2007). A Calculus for New Words. *Dictionaries: Journal of the Dictionary Society of North America*, 28, 132-138.

Cannon, Garland. (1987). *Historical Change and English Word-formation: Recent Vocabulary*. Peter Lang.

Chen, Yuzhen. (2009). Review of *A New English-Chinese Dictionary* (4th edition). *International Journal of Lexicography*, 23(2), 236-242.

Creamer, Thomas. (1995). Principles of Selection of Neologisms for a Bilingual Dictionary (English-Chinese). *Dictionaries; Journal of the Dictionary Society of North America*, 16, 102-108.

Ding, Zhicong. (2001). On the entry-selection criteria of *A New English-Chinese Dictionary*. Journal of Quanzhou Normal College, 1, 117-118.

Landau, Sidney. (1989). Dictionaries: *The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Metcalf, Allan. (2002). *Predicting New Words: The Secrets of Their Success*. Boston: Houghton Mifflin Company.

Sheidlower, Jesse T. (1995). Principles for the Inclusion of New Words in College Dictionaries. *Dictionaries: Journal of the Dictionary Society of North America*, 16, 32-44.

Wang, Fufang. (2010). Advantages and Shortcomings: A comparative study of *A New English-Chinese Dictionary* (4th edition) and *Longman Dictionary of Contemporary English* (fifth edition). *Lexicographical Studies*, 5, 74-81.

Wu, Ying. (2000). *A New English-Chinese Dictionary, Century Edition*. Shanghai: Shanghai Yiwen Publishing House.

Wu, Ying. (2001). On the century edition of *A New English-Chinese Dictionary*. *Lexicographical Studies*, 2, 119-122.

Xu, Hai. (2011). Strategies in dictionary revision—a comparative study of the third and fourth editions of *A New English-Chinese Dictionary*. *Fudan Forum on Foreign Languages and Literature*, spring, 86-91.

The Effects of Dictionary Use on L2 Error Correction

Yoshiho Satake
Surugadai University
satake.yoshiho@surugadai.ac.jp

Abstract

While some aspects of dictionary use have been illuminated by previous studies, there are still many factors and variables that have not been considered (Hartmann, 2001). To judge what the effects of dictionary use are, more empirical studies are needed. This study investigates the effects of factors and variables of dictionary use on error correction in L2 writing.

Participants' dictionary use was compared with their corpus use along with their non-use of both. The following procedure was used. In (1) the timed essay task (25 minutes), 55 Japanese intermediate EFL learners wrote an essay on a topic given by the author without consulting dictionaries or a corpus. Then, in (2) the revision session (15 minutes), the author or peer students gave feedback on errors and the participants corrected the errors, consulting English-Japanese dictionaries of their choice and the Corpus of Contemporary American English (COCA). This procedure was repeated almost every week (approximately 10 times). (3) The author collected the participants' essays and created error-annotated corpora to analyze the effects of different types of references.

The results revealed that dictionary use contributed to the correction of lexical errors with the highest accuracy. Explicit information on word usage and meaning, sometimes combined with learners' vocabulary knowledge, worked well. However, dictionary use was not very effective in correcting omission errors because there are much fewer example sentences in dictionaries than in the corpus. The findings suggest that effective dictionary use for error correction requires teachers to consider learners' L2 proficiency and teach the error types to correct with dictionary use.

Keywords: Dictionary use; error correction; error identification; L2 writing

1. Introduction

Because of great interest in the lexicon in linguistics and in the lexical aspects of second-language acquisition (SLA) teaching and learning, dictionary use has been studied (Tono, 2001). There has been a need for observations of dictionary use because the perspectives of dictionary users help researchers to examine reference needs and skills (Hartmann, 2001). Yet while some aspects of dictionary use have been illuminated by previous studies, there are still many factors and variables that have not been considered (Hartmann, 2001). To judge what the effects of dictionary use are, more empirical studies are needed. This study deals with the effects of factors and variables that dictionary use has on error correction in L2 writing.

2. Literature

A pioneering work by Barnhart (1962) found that students ranked six information categories in dictionaries as (1) meaning, (2) spelling, (3) pronunciation, (4) synonyms, (5) usage notes, and (6) etymology. This indicates that learners seem to need meaning and spelling information more than grammatical knowledge (Béjoint, 1981) and etymology (Hartmann, 2001). Moreover, Bishop (1998) writes that learners use dictionaries for various kinds of information, such as definitions, synonyms, and registers, related to the meanings of words. Thus, we may say that meaning information in dictionaries is important for learners.

The author's comparative studies of dictionary and corpus use illuminate some strengths of dictionary use for L2 learning. Satake (2015) describes it was more effective to use a dictionary when the target phrases were in both the dictionary and the corpus because consulting a dictionary was easier than consulting a corpus, as the students had generally used dictionaries before but did not have previous experience using a corpus. In addition, Satake (2014a, 2014b) states that dictionary use is more time-saving than corpus use because dictionary users look up more collocations than corpus users. Therefore, dictionary use is a good option for searching and memorizing collocations in limited time.

While some aspects of dictionary use have been illuminated by previous studies, many things remain unclear because (1) studies of dictionary users are not numerous enough, (2) the number of participants is small, (3) in most cases, only general dictionaries are studied, (4) it is difficult to evaluate and compare studies on dictionary use because the methods vary widely, (5) it is difficult to generalize the results of different studies, (6) there are still many factors and variables that have not been considered, and (7) most studies include users of existing dictionaries, not new dictionaries (Hartmann, 2001). Hartmann (2001) also states that dictionaries are effective for a wide variety of tasks like translation and vocabulary acquisition, although there is no agreement on how to evaluate their respective priorities. In addition, it is not clear how dictionaries should fit into the pattern of other teaching aids, learning strategies, and syllabuses (Hartmann, 2001). To judge what the effects of dictionary use are, more empirical studies are needed. Of the above issues, this study deals with (6), the effects of factors and variables that dictionary use has on improving L2 writing.

3. Research Questions

This study examines the effects of dictionary use on error correction in L2 writing. The research questions are as follows:

(1) What effects does dictionary use have on error correction in L2 writing?
(2) What effects does dictionary use have on error identification in L2 writing?

These questions are investigated by comparing the results of dictionary use with the results of corpus use.

4. Method

Participants' dictionary use was compared with their corpus use along with their non-use of both.

4.1 Participants

The participants in this study were 55 Japanese EFL learners. Before starting the study, the author asked the students for permission to use their essays and data for the research. On average, they reached A2 to B1 in the Common European Framework of Reference for Languages (CEFR).

4.2 Reference resources

The author allowed the students to use any dictionaries they liked because imposing only one dictionary for all the students was impractical, considering the many kinds of dictionaries they had. Almost all the students used English-Japanese dictionaries, which is reasonable to expect since foreign language learners tend to use bilingual dictionaries irrespective of their language proficiency levels (Piotrowski, 1989). Approximately two-thirds of the students used the *Genius English-Japanese Dictionary* (Konishi & Minamide, 2006) containing about 96,000 words, which has been the best-selling learner's English-Japanese dictionary in Japan for more than 25 years (Taishukan, 2014), and approximately one-third of the students used the online *Weblio English-Japanese Dictionary*, which contains nearly five million words (2016). Almost all of the students used electronic dictionaries, and indeed only one student in 2014 used a paper dictionary.

The corpus used as a reference resource was the Corpus of Contemporary American English (COCA), which is a large balanced corpus with 560 million words (Davies, 2017-). Before the first revision session, the students were given twenty minutes' instruction on using the corpus to correct errors. The students were instructed to search for the target word and interpret the concordance lines to identify which word(s) should or should not be used in the context.

4.3 Tasks

The participants completed timed essay tasks. For the first task, the participants were asked to write an essay based on the topic given by the author. The first task was timed for 25 minutes, and no access to reference resources was allowed.

In the second task, the students were given 15 minutes to correct the errors for which they were given feedback, using reference resources. For the revision task, the students were given a revision sheet on which they recorded their errors and the corrections to their essays, the reference resources they used, parts of the example sentences that they consulted for their correction, and the reference resources their peer students had employed to identify the errors.

4.4 Feedback on errors

The participants were alternately provided teacher and peer feedback. Both the author and peer students gave feedback by highlighting errors in each essay. No explanation was provided of why the highlighted words or phrases were problematic.

4.5 Error-annotated learner corpora

The author created error-annotated corpora from the participants' essays to analyze how the participants corrected their errors both quantitatively and qualitatively. The author modified and used the error tags in the NICT Japanese Learner English (JLE) Corpus, which contain the information on parts of speech and error types (Izumi, Uchimoto, & Isahara, 2004).

4.6 Procedures

The following procedure was used:

(1) The first timed essay task was given (25 minutes in class).

(2) After the first task, a revision session was held in which the students were given feedback on errors of their original essays. The students were given alternate teacher feedback and peer feedback.

(3) The students undertook a revision session for 15 minutes, consulting reference resources.

The above procedure was repeated approximately ten times.

(4) The learner corpora of the participants' essays with error annotation were created.

(5) The author conducted error analysis based on the error annotations.

5. Results and Discussion

5.1 Effects of dictionary use on error correction

5.1.1 Error types that were corrected with dictionary use

The participants corrected 349 errors with dictionary use, and the average rate of accurate correction was 73.9 percent. Table 1 shows the eight most common error types that the participants corrected more than 10 times with dictionary use, which were, in order, lexical errors, omission errors, number errors, part-of-speech errors, spelling errors, addition errors, verb form errors, and tense errors.

To judge whether there was a significant difference in the frequencies of accurate and inaccurate corrections of errors among the different reference resources, the author used a chi-squared test. As for lexical errors, the author found a significant difference ($\chi^2(2) = 15.24$, $p < 0.01$, Cramer's $V = .24$), and residual analysis showed a significant difference between dictionary use and no use of reference ($\chi^2(1) = 15.24$, $p < 0.01$, Cramer's $V = .21$). Since the rate of accurate correction with dictionary use was higher than that with no use of reference, we can say that dictionary use promoted significantly more frequent accurate corrections of lexical errors than no use of reference resources. There was no significant difference between the effects of dictionary use and those of corpus use.

As for number errors, part-of-speech errors, spelling errors, addition errors, tense errors, and verb form errors, there was no significant difference between the two reference material types.

Concerning omission errors, the author found a significant difference ($\chi^2(2) = 10.22$, $p < 0.01$, Cramer's $V = .18$), and residual analysis showed a significant difference between dictionary use and corpus use ($\chi^2(1) = 6.46$, $p < 0.05$, Cramer's $V = .14$). Since the rate of accurate correction with dictionary use was lower than that with corpus use, we can say that dictionary use promoted significantly less frequent accurate corrections of omission errors than corpus use. There was no significant difference between the effects of dictionary use and those of no use of reference.

In short, the results suggest that the strength of dictionary use was more accurate correction of lexical errors in comparison to no use of reference, and its weakness was less accurate correction of omission errors than corpus use.

	Dictionaries			Corpus			No use of reference		
Error types	Number of corrections	Number of accurate corrections	Rate of accurate correction	Number of corrections	Number of accurate corrections	Rate of accurate correction	Number of corrections	Number of accurate corrections	Rate of accurate correction
lexical	92	69	75.0%	77	55	71.43 %	107	54	50.47 %
omission	74	51	68.9%	137	115	83.94 %	116	79	68.10 %
number	34	30	88.2%	40	34	85.00 %	100	77	77.00 %
part-of-speech	22	16	72.7%	13	11	84.62 %	27	18	66.67 %
spelling	17	17	100.0 %	5	4	80.00 %	26	20	76.92 %
addition	12	11	91.7%	34	28	82.35 %	33	30	90.91 %
tense	12	11	91.7%	4	3	75.00 %	63	57	90.48 %
verb form	12	7	58.3%	5	3	60.00 %	21	19	90.48 %

Table 1: The eight most common error types that the participants corrected with dictionary use and the comparison with corpus use and no use of reference

5.1.2 Strengths of dictionary use for error correction in L2

To explain the strengths of dictionary use, the following three examples of lexical errors that the participants corrected with dictionary use are presented. The errors identified by the author or a peer student are underlined.

(a) So many people worried by stomach cancer.
(b) The two women are grateful people.
(c) She always became the sample.

In (a), the participant looked up the word "worried," cited the example sentence "We worried about whether the lecturer would arrive in time," and used this information to correct his sentence. It would have been easy for the participant to understand his expression was wrong because *Genius*, the dictionary he used, gives an explicit explanation of word usage: it explains that "about," "for," or "over" follows the intransitive verb "worry." In addition, dictionary use was helpful because the participant found example sentences that included the phrase "worried about."

As for (b), the participant looked up "grateful" first and "remarkable" second, cited "a person of remarkable ability" as an example, and used this information to correct her sentence. It would have been easy for the participant to understand her expression was wrong because the definition of "grateful" was not the meaning she intended. It seems she wanted to use "great," but she did not look up the word, probably because she did not know the correct spelling of the word. She then looked up "remarkable," which is a synonym of "great," and decided to use that word for the correction. The information on the meanings of the words helped her decide which word to use for the correction. Compared to (a), it would have been difficult to correct her error because she needed to know the word with the meaning she wanted to express in order to consult the dictionary.

In the case of (c), the student looked up "sample" first and "model" second and cited the sentence "She is a model of honesty" as the information used to correct her sentence. It would have been easy for her to understand her expression was wrong because the definition of "sample" was not what she intended. Then she looked up "model," the definition of which is what she intended, and decided to use that word for correction. As with (b), the information on the meanings of words helped her decide which word she should use for correction. As with (b), it would have been more difficult to correct her error than (a) because she needed to know the word with the meaning she wanted to express in order to consult the dictionary.

Although (a), (b), and (c) belong to the same error category "lexical," they are not exactly the same type of error. In the case of (a), the combination of the verb "worry" and a preposition is important in choosing the appropriate word. When a combination is often used, it is likely that the dictionary contains it, and it should be easy to correct such errors with dictionary use. In the cases of (b) and (c), knowing the word with the intended meaning is important in choosing the appropriate word. When learners have enough vocabulary knowledge to use another word with a similar meaning, meaning information in dictionaries helps, and it would in that case be easy to correct errors with dictionary use. However, when learners do not have sufficient vocabulary knowledge and do not know another word with a similar meaning, they cannot look up the appropriate option and thus cannot correct their errors with dictionary use. Therefore, both the meaning information of words and learners' vocabulary knowledge influence the correction of meaning-related errors with dictionary use.

Thus, we see that explicit information on word usage and meaning, sometimes combined with learners' vocabulary knowledge, helped the participants correct their lexical errors.

5.1.3 Weaknesses of dictionary use for error correction in L2

Example (d) contains an omission error, which one of the participants attempted to correct with dictionary use but could not correct accurately.

(d) At that time it was unbelievable because Nobunaga was <u>son</u> of a shogun so his social position is different from them.

The participant looked up the word "son," cited the example phrase "the sons of Adam," and used this information to write "the sons of a shogun," a wrong attempted correction. The student chose and referred to an inappropriate example because his vocabulary and grammatical knowledge were insufficient. That is, without enough vocabulary and grammatical knowledge, he could not use the dictionary information appropriately. The participants read much fewer example sentences with dictionary use than with corpus use

and could not access information on the frequency of co-occurrence words, which could have led to the lower rate of accurate correction of omission errors with dictionary use.

5.2 Effects of dictionary use on error identification

By using dictionaries, the participants identified 19 lexical errors. The lexical error was the only error type for which the participants identified more than nine errors with dictionary use. Since lexical errors were included in the top eight error types that the participants corrected by using dictionaries, we may assume that the same strength of dictionaries mentioned above (i.e., providing information on the meanings of the target phrases), helped them identify these errors, as it helped them correct their own lexical errors. The following are examples of lexical errors that the participants identified with dictionary use. The underlined word(s) are the sections that the participants identified.

(e) He is left style, I have never seen his lose in armletheling.
(f) So many people worried by stomach cancer.

In (e), it would have been easy for the participant to identify "left style" as an error once he consulted a dictionary because he could not have found this expression in the dictionary, while he might not have found correct expressions like "left-handed" or "a left-hander." As for (f), once the participant looked up "worry," he would have found some combinations of "worry" and a preposition and discovered that this expression was wrong. Compared to (e), the case of (f) seems more difficult because the phrase "worried by" is included in dictionaries. The participant would have needed to analyze some example sentences to judge whether the expression was appropriate in the context or not. The strength of dictionaries, which is that they provide meaning information, could have helped the participant when he examined example sentences. To use the meaning information appropriately to make an accurate judgment for error identification, the learner would have needed some knowledge about the usage of "worry." We can say that the participant needed more vocabulary knowledge to identify the error in the case of (f) compared to (e). In short, the results show that dictionaries' meaning information aided the participants' identification of lexical errors.

6. Conclusion

This study uncovered the effects of dictionary use on L2 error correction. The results can be summarized in the following two areas: (1) the effects of dictionary use on error correction in L2 writing and (2) the effects of dictionary use on error identification. As for (1), dictionary use was effective in promoting the accurate correction of lexical errors, as it provides meaning information, which helps because lexical errors are related to word meaning. Explicit information on word usage and meaning, sometimes combined with learners' vocabulary knowledge, worked well. However, dictionary use was not very effective in correcting omission errors when the participants could not appropriately use dictionary information due to lack of vocabulary or grammatical knowledge. Another weakness is that are much fewer example sentences in dictionaries than in the corpus, and thus the participants could not use frequency information on the co-occurrence of words, unlike the corpus. Regarding (2), as well as error correction, dictionary use was useful to identify lexical errors, as meaning information helped.

The results lead to the conclusion that teachers should take error types into account in order to effectively use dictionaries for accurate error correction and identification because it is suggested that the effectiveness of dictionary use for error correction depends on error

types. The findings also suggest that effective dictionary use for error correction requires teachers to consider learners' L2 proficiency because some vocabulary knowledge was needed to use the meaning information in dictionaries effectively. The pedagogical implication of the research is that, to maximize the strength of dictionary use for error correction in L2 classrooms, instruction on dictionary use is needed: teachers should consider learners' L2 proficiency and provide learners with instruction on the error types for which dictionary use is effective and ineffective.

References

Barnhart, C. L. (1962). Problems in editing commercial monolingual dictionaries. In F. W.

Davies, M. (2017-). *The Corpus of Contemporary American English*. http://corpus.byu.edu/coca/

Householder, & S. Saporta (Eds.), *Problems in Lexicography* (pp. 161-181). Bloomington: Indiana University Press.

Bishop, Graham. 1998. Research into the use being made of bilingual dictionaries by language learners. *The Language Learning Journal*, 18(1): 3-8. doi: 10.1080/09571739885200181

Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: a study of language needs and reference skills. *Applied Linguistics, 2*, 207-222.

Hartmann, R. (2001). *Teaching and Researching Lexicography.* Harlow: Longman.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). *Nihonjin 1200nin no Eigo Speaking Corpus.* Tokyo: Alc.

Konishi, T., & Minamide, K. (Eds.). (2006). *Genius English-Japanese Dictionary* (4th ed.). Tokyo: Taishukan.

Piotrowski, T. (1989). Monolingual and bilingual dictionaries: fundamental differences. In M. Tickoo (Ed.), *Learners' Dictionaries: State of the Art* (pp. 72-83). Singapore: SEAMEO Regional Language Centre.

Satake, Y. (2014a). *Corpora vs. dictionaries: their effects on learning English collocations in L2 writing tasks*. Presented at the Second Asia Pacific Corpus Linguistics Conference (APCLC). Hong Kong, March 7-9.

Satake, Y. (2014b). How does a corpus influence learning L2 collocations? *Teaching and Language Corpora: Eleventh International Conference (TaLC 11)* (pp. 107-108). Lancaster: Lancaster University.

Satake, Y. (2015). Comparison of Dictionary use and corpus use: Different effects on learning L2 phrases. In *Proceedings of ASIALEX 2015 Hong Kong*, eds. L. Li, J. Mckeown and L. Liu, 222-228. Hong Kong: Hong Kong PolyU.

Taishukan. (2014, 12 16). Genius Eiwa Jjiten Dai-Go-Han kanko. Retrieved 8 13, 2016, from PR TIMES: http://prtimes.jp/main/html/rd/p/000000001.000012210.html

Tono, Y. (2001). Research on Dictionary Use in the Context of Foreign Langugae Learning. Tübingen: Max Niemeyer Verlag.

Weblio English-Japanese Dictionary and Japanese-English Dictionary. (2016). http://ejje.weblio.jp/

Contact Us

King Mongkut's Institute of Technology Ladkrabang

Chalongkrung Rd. Ladkrabang, Bangkok Thailand 10520

asialex2018@kmitl.ac.th

+66 (0) 2329 8445